Interpreting Idea Maps: Pairwise comparisons reveal what makes ideas novel

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 to score ideas on their uniqueness. However, their decisions about 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 participants 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—colored polygons (known answer) and milk frother sketches (unknown answer). We show that idea maps shed light on factors considered by participants in judging idea similarity and the maps are robust to noisy ratings. We also compare physical maps made by participants on a white-board to their computationally generated idea maps to compare how people think about spatial arrangement of design items. This method provides a new direction of research into deriving ground truth novelty metrics by combining human judgments and computational methods. Open-Source code implementing our approach is available at: `https://github.com/IDEALLAB/idea_map`

1 Introduction

Creativity is the driving force of innovation in the design industry. Despite many methods to help designers enhance creativity 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 to decide how to judge all the submitted ideas. Contributors have just sent in a flood of design ideas of variable quality and these ideas must now be evaluated 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 attempts to quantitatively assess creativity of ideas [3–5]. Evaluating creativity is often viewed as the comparison of design ideas for utility, novelty and surprise [6]. Utility or quality is a measure of the designs' performance [7] and can depend on 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 [8]. An idea may be considered surprising because it is unlikely or unfamiliar. When a new idea unexpectedly falls into an already familiar conceptual framework (or thinking style) one is intrigued to not have realized it before. The focus of this paper is novelty measurement from pairwise comparisons asked from people.

Novelty measurement is key to many research and industrial endeavors, for example in patent decisions. To accept a patent application, it "must be new, that is, bestowed for the first time upon the public by the patentee" [9] or in other words, it should have a high novelty score compared to existing items in the market. Novelty of ideas is usually measured by human judges, automated methods or a combination of the two. When ideas are being judged by humans, the judges are often experts in the domain with substantial knowledge of the field and of the market. They can thus provide more informed and trustworthy evaluations [10]. Many crowdsourcing platforms such as Topcoder, Taskcn, and Wooshii use expert panels to select contest winners [11]. 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 also been proposed [12] to evaluate ideas due to their large diversity of viewpoints, knowledge and skills [13]. Crowdsourcing has been used in an array of studies for evaluation of items. For instance, Yu *et al.* [14] used crowdsourcing to guess the meaning of 20 US Pharmacopeial Convention pictograms. To obtain crowd evaluation similar to expert panel, Wu *et al.* [15] recommended that before crowdsourced workers are used to evaluate designs, one should collect the evaluation criteria from the crowd itself, and then use those crowdsourced evaluation criteria to evaluate designs. However, there is no clear evidence demonstrating that crowds can be used as a proxy for experts' evaluations to assess a large number of complex ideas [16]. In this paper, we focus on uncovering factors in subjective novelty ratings of participants, irrespective of them being 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 for novelty ratings?" In this paper, we address these questions by calculating what we call *idea maps*, *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 related work in creativity ratings and design visualization. Thereafter, we provide 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.

2 Related Work

In this section, we review research related to creativity ratings and design space visualization, which relate directly to our work in creating explainable metrics.

2.1 Creativity Ratings

In the social sciences, creativity is often measured subjectively through the Consensual Assessment Technique (CAT) [17]. 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 [18]. 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, 19, 20]. For example, one of the commonly used tree-based metrics [21] breaks down creativity into quantity, quality, novelty, and variety. These methods are widely adopted in engineering due to limited participant bias [22]. The resultant novelty score of an idea depends on which attributes are considered in the tree and may vary between two different participants or trees [23]. 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 [24, 25]. In this paper, we propose a third approach, which combines the strengths of both methods by asking simple subjective queries from participants and then using computational methods to estimate idea novelty. In the

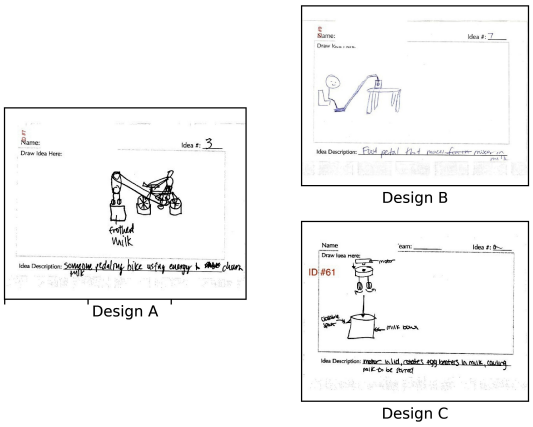

Fig. 1. Example of triplet query asked from participants in our experiment. Participant answers the question: "Which design is more similar to design A?"

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 participants is to visualize the design space by placing all ideas on a map and grouping similar items together. Design space exploration techniques [26] were 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 [27–29]. A difficulty in creating low dimensional manifolds for design ideas is that complex design ideas often lack compact vector representations or known similarity measures. To circumvent this problem, one can directly ask people 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) [30]. Methods like classical multi-dimensional scaling [31] can be used with these ratings to find an embedding. However, these ratings are not considered suitable for human similarity judgments as different participants use different "internal scales" and participants may be inconsistent in their grading [32].

As humans are better at comparing items than giving absolute scores [33], 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) [34], Crowd Kernel Learning [35] and Stochastic Triplet Embedding [32]. Such methods take triplet ratings as input and output either an embedding or a similarity kernel between items which best satisfy triplet responses from one or more participants.

Techniques for capturing similarity among items using triplets have been applied in many areas like computer vision [36], sensor localization [37], nearest neighbor search [38] and density estimation [39]. In [40], 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.* [41] use triplet similarity comparisons by crowdworkers to create spatial idea maps. They show that participants 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 participants. Methods with output similar to idea maps are often used in brainstorming sessions by UX designers. For instance, affinity diagrams [42] were introduced in the 1960s, alongside the KJ Technique by a Japanese anthropologist. In affinity mapping, large amounts of language data (ideas, opinions, issues) is collected and organized into groupings based on their relationships. However, unlike idea maps, the Euclidean distance between each item within a group does not hold any meaning in an affinity map, nor do the distances between groups. Hence, these maps are not suitable for assessing novelty of design items.

3 Methodology

Our methodology relies on asking simple subjective queries (called triplet queries) from participants and finding idea maps, which best satisfy the responses to those queries. 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

To generate idea map for a set of designs, we ask triplet queries from participants about those designs. Figure 1 shows an example of a triplet query with three design sketches used in our study. The participant answers the question: "Which design is more similar to design A?", giving a subjective assessment of how similar two designs are. We generate all possible triplet queries from a given set of N designs and give those queries to participants as surveys. After collecting responses to these queries, we use the Generalized Non-metric Multidimensional Scaling (GNMDS) technique [34] to find embeddings of design ideas.[1] The idea map obtained by applying GNMDS to the triplet responses by a participant 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 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. We code the response of the participant to any query (*e.g.*Figure 1) 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 of GNMDS is x,y coordinates for each design item, which can be plotted to give the idea map.

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 [8], 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:

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

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

X_i, Y_i are the 2-D coordinates of an idea i and 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 [43]. Nov_{cent} defines the novelty of an idea as the distance from the centroid and has been used in [7] 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 easy to compute and make few assumptions about the distribution of ideas or how ideas are clustered in the map. It is possible to compare 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 for few known metrics. Finding the best novelty metric (depending on how one defines best metric and if assuming such a metric would generalize to all domains) for any given domain is outside the scope of this paper.

3.3 Measuring Participant Performance

Triplet responses given by participants can vary in accuracy or reliability due to factors like expertise or motivation. However, it is difficult to assess the quality of triplets by measuring intra-rater reliability, as they are subjective assessment

[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.

of how a participant views the similarity of ideas. Instead, we estimate a participant's performance by measuring how consistent they are with their own responses using two methods. First, we estimate the self consistency of participants by adding additional triplet queries, which are repeats of existing queries. Second, we measure the number of violations a participant makes in the transitive property of inequality; for example, suppose a participant 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 participant provides a third triplet response of *BCA* indicating idea B is more similar to idea C, then this violates the 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 participant performance. In this study, we do not use explicit criteria to filter out participants with lower scores on these metrics but this information could be incorporated in future studies to give more importance to idea maps of participants who are more self-consistent.

3.4 Measuring Map Similarity

To judge how maps differ between different people or using different methods, we need to define measures which can quantify the difference between two maps. To find similarity between two idea maps, we use following three error methods:

Disparity error The disparity score measures how different two maps are if one is overlaid on another. To compare the 2-D position of points on two maps, they should be on the same scale and corresponding points should align. As maps can be on different scales and may be rotated or translated, we first correct for these variations using Procrustes analysis [44]. It finds the optimal scaling, rotations, and reflections such that the sum of the squares of the point-wise differences between the two input collection of points is minimized. The least squared error (after transformation of one map) is called the 'Disparity' score between the two maps.[2] As the disparity score depends on an intermediary step of correctly solving another optimization problem, this 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.

Distance error In this method, instead of comparing the positions of maps, we measure the corresponding intra-point distances between the two maps. To do so, we first calculate the Euclidean distance vector of each point in a map with every other point in the same map. For 10 points, we get 45 unique distances. Next, we find the mean squared error (MSE) between the distance vectors of the two maps. As distances are rotation and translation invariant but not scale 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.

Triplet error The above methods measure how metric distances between two maps differ. As the maps obtained using triplets are non-metric, it is possible that two maps with significantly different spatial arrangements satisfy the same set of triplets (*i.e.*, are equally consistent with human-provided triplets). To account for this, we define another score which depends only on relative distances. 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 participants). We count the number of triplet responses in common between the two maps. Triplet error is defined as the percentage overlap between these two sets of triplets *i.e.* $\frac{|S_1 \cap S_2|}{|S_1|}$. It measures how the two maps compare in relative distances of items.

To explain a case where triplet error maybe is zero while other measures maybe high, we take the example of comparing two maps with four items each shown in Fig. 2b. Visually the two maps look different and even after scale and rotation transformations, the points do not overlap. However, if we list triplet responses which satisfy the map on left side, we get the following set of twelve triplets S_1=[ABC, ABD, ACD, BAD, BAC, BDC, CAB, CDA, CDB, DBA, DCA, DBC]. As mentioned before, triplet ABC means A is closer to B than C. Now, if we list the triplets satisfying the map on the right side, we find that all triplets in set S_1 are satisfied in the map. Hence, the triplet error is zero between these two maps. We report all three measures when comparing maps in our experiment section.

4 Experimental Results

To demonstrate our methodology, we consider two case studies. We chose the first case study such that the generated idea maps are simple to understand and the novelty measure is easily verifiable. By selecting items with only a few attributes,

---

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

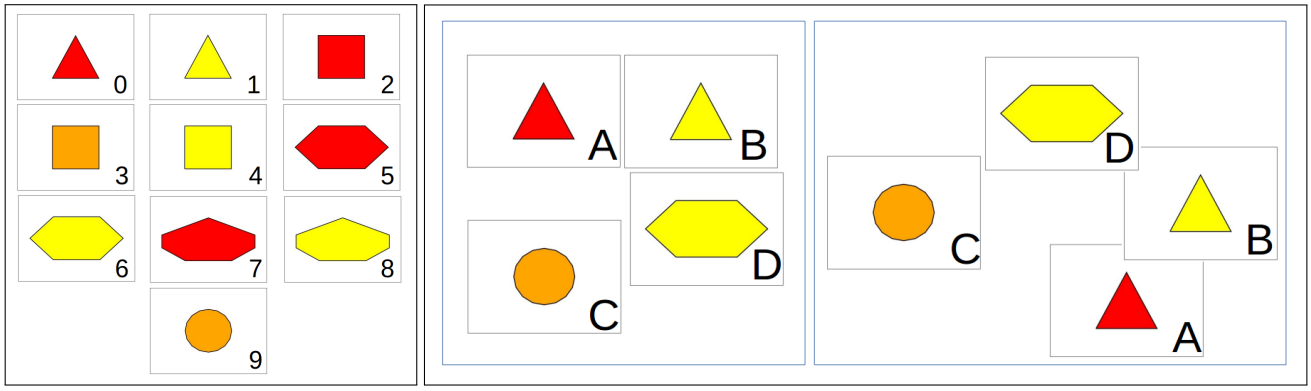

Fig. 2. a) Dataset of ten polygons used in first experiment. b) Two idea maps with 4 items each. Although these maps look different, they satisfy the same set of triplet queries.

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 participants 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 participants. We show two dimensional idea maps and novel items discovered for different participants in a seemingly simple design domain. In the second study, we selected ten milk frother sketches from an 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. We also ask participants 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. 2a, 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 participants. The participants 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, participants have to decide whether they find the red heptagon more similar to the yellow hexagon or the red square. One participant 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 participants to explain their choice for 20 randomly selected triplets. These responses helped to verify our hypotheses about the factors considered by each participant.

Map obtained using fixed rules To verify that the triplet generated maps correctly reflect the triplet responses provided, we first define fixed set of rules by which the triplets are answered. We call this set of rules *the automated rater*. We define the rules such that polygons with the least difference in number of sides are always rated as most similar—*e.g.*, a triangle will be rated closer to a square than a hexagon. When a polygon B and polygon C have similar priority in a previous rule compare to polygon A, the automated rater selects the polygon which is more similar in color to the polygon A. For example, if we compare a yellow triangle and a red pentagon with their similarity to a red square, the red pentagon is considered closer.

As the automated rater uses consistent rules in judging all triplets, it has a self consistency score of 100% and has zero transitive violations. The resultant idea map obtained from the automated raters triplet ratings is shown in Fig. 3. This idea map shows that similarly shaped items group together. As one might expect, the two dimensions identified by 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 closer compared to the gap between squares and hexagons. This is because triplets with smaller difference in their number of sides are rated as more similar by the automated rater. Hence, we can verify that the map obtained is a good representation of the triplet ratings provided by the automated rater.

Maps obtained using human participants In contrast to the automated rater, human participants may not always use consistent rules. Different people may give different priority to attributes like color, shape, symmetry, *etc.* We summarize our results for 11 participants in Table 1. Column 2 lists the self consistency score for each participant and column 5 lists the count of transitive violations. Columns 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).

Take the example of the idea maps obtained for two participants (participant id 5 and participant id 9 from Table 1 respectively). Participant 5's idea map, shown in Fig. 4, places similar shaped polygons near to each other and the red colored polygons are placed above yellow ones. This provides evidence that this participant used shape and color as main criteria to answer triplet queries. In contrast, participant 9's idea map placed similarly colored items together (Fig. 5), indicating that similarity in color matters more 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 participant 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 participant 9 mentions many of her triplet comparisons were decided based on "color, shape, number of sides". Hence, the criteria used by individual participants 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 participant 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 participant using the two novelty metrics. We find that the two metrics give the same set of Top 3 items for 9 participants and remaining 4 have at least 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 the top three for most participants; all novelty metrics indicate that the circle is the most novel item and there is consensus among participants that it is the most novel item in the set. This matches our expectations, as we designed this dataset such that the circle differs both in color and 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.

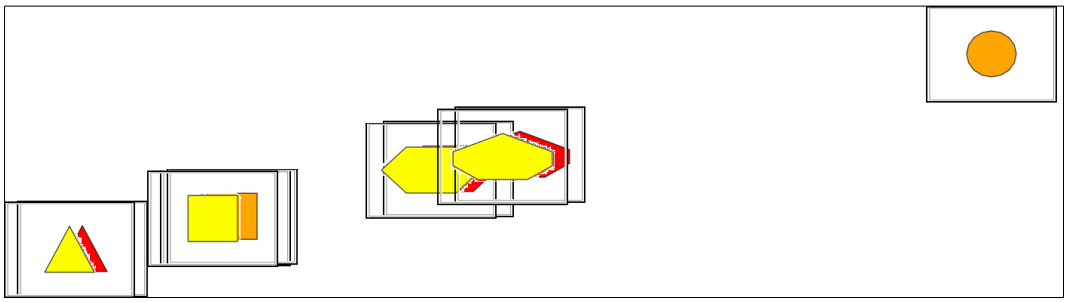

Fig. 3. 2-D embedding obtained for polygons using triplets given by the automated rater, who uses preset rules.

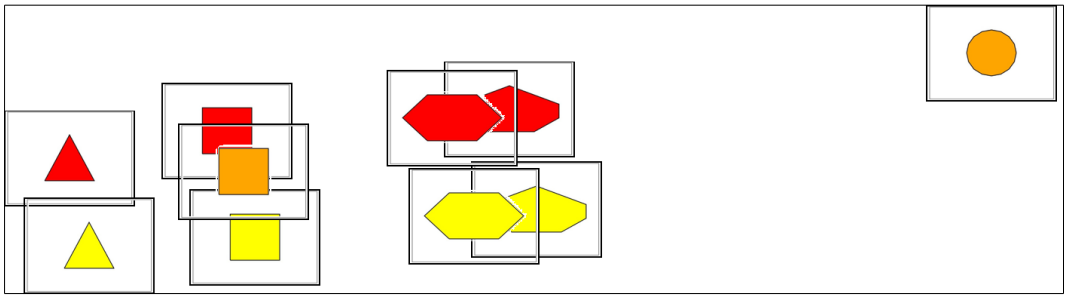

Fig. 4. 2-D embedding obtained from polygon dataset by participant 5, who uses number of sides as primary criteria the decisions

4.2 Experiment 2: Design Sketches

In this experiment, we study the embedding for electro-mechanical items—specifically, ten design sketches of milk frothers. This set of design sketches is randomly sampled from a larger dataset of milk frother sketches [1, 45]. The entire dataset is publicly available at `https://sites.psu.edu/creativitymetrics/2018/07/18/milkfrother/`. To create the original dataset, the authors recruited engineering students in same first-year introduction to engineering design

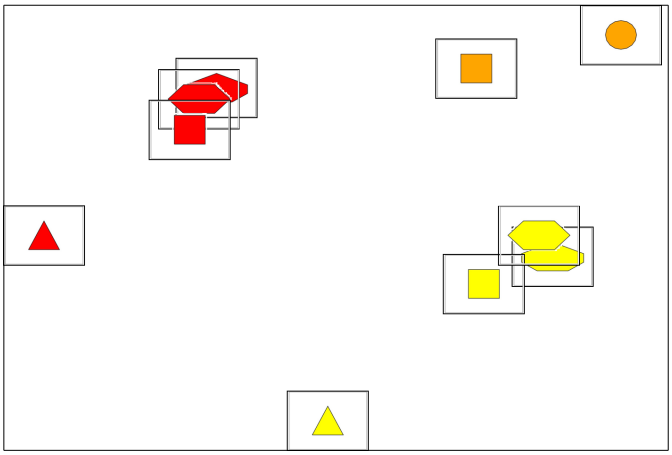

Fig. 5. Two dimensional embedding obtained from polygon dataset by participant 9, who uses 'color, shape, number of side' as criteria

Rater id	Self consis-tency (%)	Top 3 Nov_{sum}	Top 3 Nov_{cent}	Transitive viola-tions	Triplets not satis-fied (%)
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. Participant performance and Top 3 novel items for different participants of experiment 1 on polygons. We find that most participants find circle (item 9) as the most novel polygon

Rater id	Self consis-tency (%)	Top 3 Nov_{sum}	Top 3 Nov_{cent}	Transitive viola-tions	Triplets not satis-fied (%)
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. Participant performance and top three novel items for different participant of experiment 2 on design sketches. We notice higher variance between participants as well as lower self-consistency.

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 the experimental procedure for data collection are available online.[3]

We obtained ten random samples (without replacement) from this dataset for this experiment. Figure 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 the idea maps instead. Similar to the previous case, eleven participants were used in this experiment. The participants comprised one professor (Industrial Engineering), two Ph.D. students (Industrial Engineering) and seven under-graduate students (Psychology). Two participants (rater id 5 and 10) are considered experts in rating milk

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

frother sketches. To qualify expertise, both experts has at least 4 years of applied experience in design and assessment and had published at least four papers in design and creativity assessment.

Figure 7a and 7b show the idea maps obtained by participants 7 and 10. These maps provide useful cues into the decision making process of these participants, who used different decision making criteria. The embedding of participant 7 in Fig. 7a provides evidence that she might have grouped sketches which have a 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, participant 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 participants, we qualitatively analyzed their explanations. For the triplet query shown in Fig. 1, participant 7 finds sketch C as more similar to sketch A and mentions her choice as being based on "Simple or complex" design. Participant 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 participant 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, participant 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 participant 10, ideas in her map are possibly grouped due to a combination of different factors.

To verify the novelty calculation for participant 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 the Shah, Vargas-Hernandez, Smith (SVS) metric [21] 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 participant 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 the justifications given by experts (who have significant experience in rating milk frothers) and novices: 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.

The 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 participants 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. The larger error may be caused by different factors—differences in motivation, expertise, higher complexity task, participants changing judging criteria during survey or differences in focus and mental block effect.

To understand how sketches are grouped together, we combine the triplet responses of all participants and obtain a joint idea map. Figure 8 shows the joint map of all eleven participants. As we add all triplets from participants who considered different (unknown) factors in judging idea similarity, the aggregated map represents 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 participants 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 participants.

So far, we have shown how individual idea maps can provide cues into factors important for participants in judging idea similarity. We have also shown how a joint map of multiple participants meaningfully groups sketches and can be used to estimate explainable novelty of sketches. Note that our method finds the idea map for each individual, where the idea map of an individual represents how the participant views similarity between ideas. Along with calculating maps corresponding to each individual, we aggregated the triplet responses to show the map which represented the majority opinion on how the participants viewed similarity between ideas. Hence, aggregation is not central to finding novelty of ideas and the method allows researchers to study idea maps separately as well as by aggregating it. Next, we measure how participants differ from each other in their triplet responses.

Similarity between participants To compare the similarity between triplet responses of different participants, 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

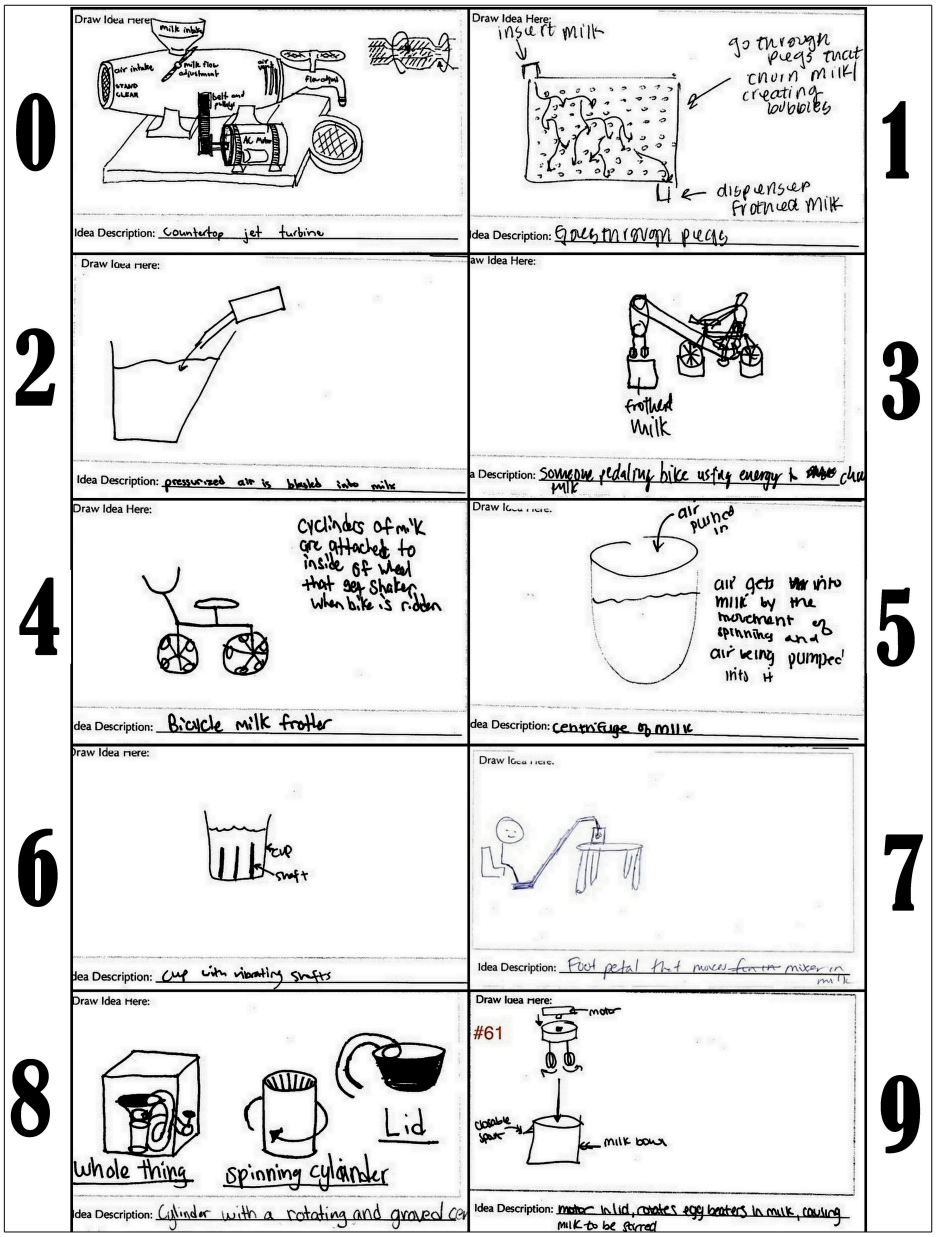

Fig. 6. Ten milk frother sketches used in experimental 2.

that participants 1, 3, 5 and 10 are in the first cluster and all other participants are in the second cluster. Interestingly, participants 5 and 10 were the two experts in our participant pool and we found that they were also clustered together, along with participant 1 and participant 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 participants 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 of group of participants.

Sketches that are difficult to judge Different sketches have different levels of complexity. Some sketches in a triplet query can be considered similar or 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 participants is important, as it can help understand features within these difficult sketches which cause disagreement among participants. 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 participants disagreed. For instance, if 50% of participants 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

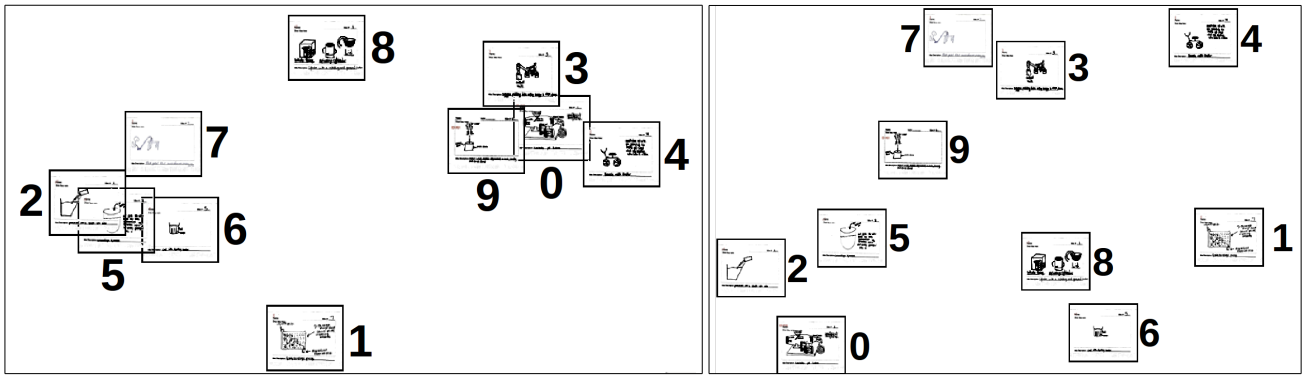

Fig. 7. Left: (a) Idea map of design sketches for participant 7. Center of the sketch represents the 2-D position of embedding. Two main clusters can be seen. Right: (b) Idea map of design sketches for participant 10. Center of the sketch represents the 2-D position of embedding.

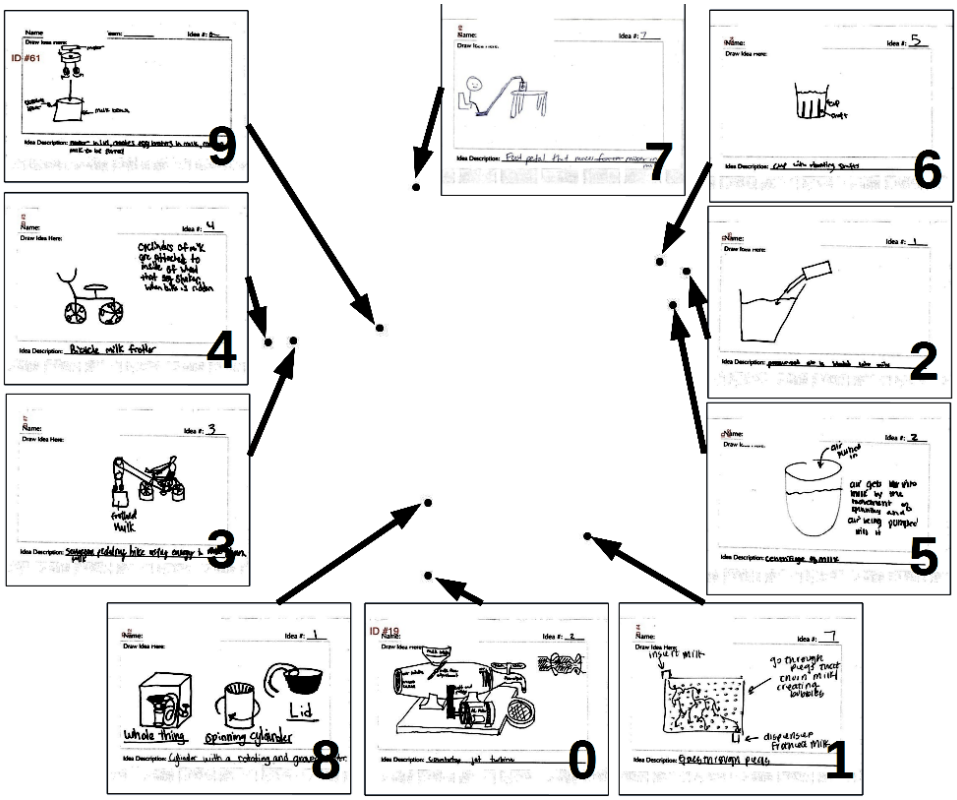

Fig. 8. Idea map obtained by combining triplets from all participants. Id of each sketch is at bottom right corner.

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 participants in which it appears. Using this score, we find that sketch 8 has the highest disagreement score among participants, followed by sketch 0. Sketch 1 followed by sketch 6 have least disagreement scores. This indicates whenever sketch 8 appeared in a triplet, participants 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 participants 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 participants is 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 participants) and show that these maps are resilient to significant levels of noise.

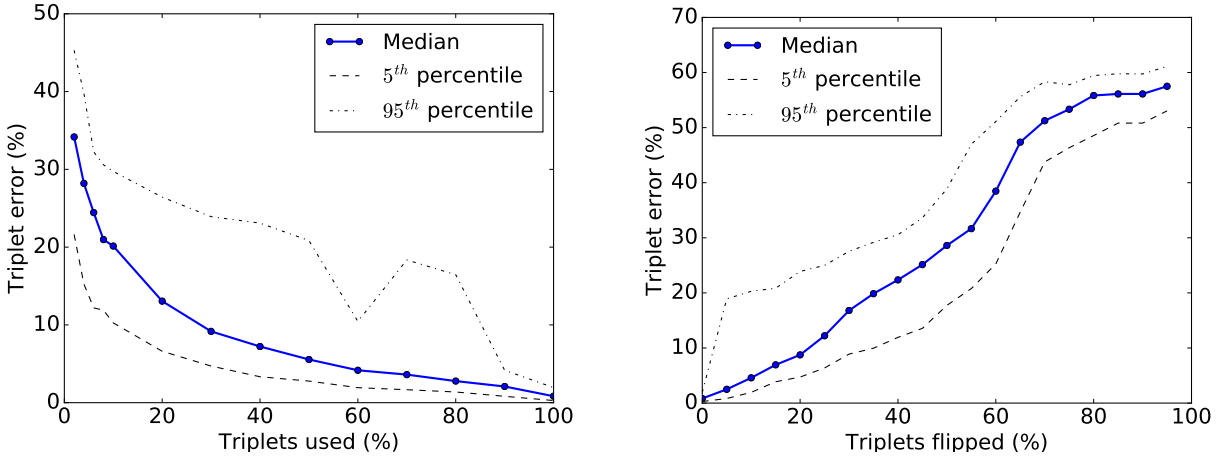

Fig. 9. Left: (a) Triplet error between idea maps of embedding shown in Fig. 8 and embedding obtained using a subset of triplet ratings. 100 runs with different subsets used to obtain embeddings. Using only 30% of triplets, median error is less than 10%. Right: (b) Triplet error between embedding generated 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 participants.

4.3 Maps using fewer triplets

We collected 360 similarity judgments each from 11 participants 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 [46]. 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 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 participants 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 participants, but randomly flip the response for a percentage of those triplets. This situation can occur in cases where participant accuracy goes down due to fatigue, when a few participants intentionally lie about similarity judgments, participant 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 participant would generate directly? (That is, 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 participants and then compare these idea maps with the maps generated using embedding methods.

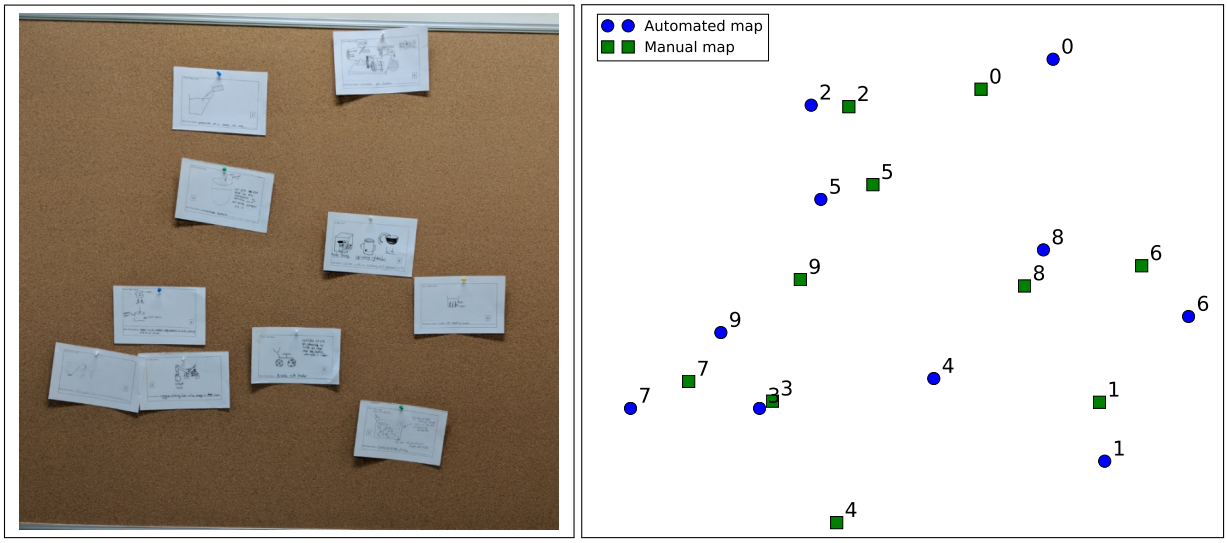

Fig. 10. Left: (a) Participant creating a map by positioning idea sketches on provided canvas. Right: (b) Correspondence between human generated map and triplet map for participant 10 after Procrustes transformation. Apart from sketch 4, most sketches have minor relative displacements.

5.1 Participants

Four subjects were selected from the group of participants 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 with the intent to reduce priming, and obtain results that are less affected by 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. As part of this research, we did not include analysis of design process used by the participants using audio and video records. Such an analysis is of independent interest and does not fit in the scope of this paper.

Figure 10(a) shows how the maps were pinned on a board by one of the participants participating in this experiment.

5.3 Comparison between automated and manual maps

We compare manual idea maps with the automated idea maps (generated using triplet embedding methods) by 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 participant 10. We notice that her automated and manual map align quite well, as seen by similar numbers (sketch id) positioned nearby each other. However, that is not the case for all participants. Table 3 summarizes the results for all four participants. The second column shows the percentage of triplets satisfied by the human generated map. It measures percentage of triplet inequalities (from the survey taken by the same person) that are satisfied by the map generated by the person. Higher values mean that the person's manual idea map differs more from their survey responses. The third column shows the triplet error between map obtained using automated method and manual map. We notice that this error increases for participants who have low self consistency (column 2). The fourth and fifth columns show the disparity measure and mean squared error, both of which have a similar trend as triplet error.

We observe that participant 10 has the highest alignment between her manual and automated maps using all three metrics, while participant 5 has the least alignment. High disparity due to lack of alignment between automated and manual maps may occur for multiple reasons. It is possible that different participants had different motivations in creating the idea maps

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 participants and their triplets and triplet embedding maps. We observe that manual maps are not great at satisfying triplets.

or taking the survey. It is difficult to determine if a participant took the task non-seriously except by making assumptions on factors like the time on task. Another factor responsible for high disparity can be that participants do not understand the instructions and the idea map created manually represents a different construct than the queries answered in the survey triplets. It is also possible that people are not good at constructing such maps which require simultaneous comparison of each item with multiple other items—an observation supported by past literature [40]. It can also be a combination of multiple factors which caused the differences between idea maps for few participants. Our future work will try to uncover the reasons behind these differences by qualitatively analyzing the decision making process of raters.

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 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 or other descriptors for the entire dataset.
2. Generating such maps can help in understanding a design domain. For instance, one can use these maps to understand what features are more important in defining similarity between ideas. We find in our experimental results that participants form identifiable clusters in idea maps. This permits new ways of finding and studying fine-grained details in how people reason about concepts and designs. One can also measure changes in the idea maps of a person or team before and after some trigger event (like showing analogies) to understand what changes his or her perception of a 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 participants 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 participants 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. When comparing embedding methods and novelty metrics, future studies should take into account robustness to noise too.
5. As shown when clustering participants, we can measure the similarity between participants from their triplet responses. This similarity measure can be used to find groups of similar participants. These groupings can be used to find aggregated maps for different groups and study differences between idea maps of a group of participants. For example, it can help unpack differences in how experts rate items compared to novices, or how groups of experts from different fields might differ. Measuring differences between participants can help in training them by understanding what features someone is not paying attention to and providing an 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, repeatable 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 for defining new families of variety and novelty metrics, which can help assess ideas. In this paper, we have used simple novelty metrics like the 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 [47] to estimate novelty. One can also use volume based coverage methods like Determinantal Point Processes (DPP) [48] 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 [49]—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. To study idea maps for larger datasets in future work, active learning and better elicitation approaches can be used to reduce the number of triplet queries. Another approach to reduce the number of triplet queries is to leverage the information available about design items like known design attributes. In such cases, it is possible to integrate existing features of the designs with human triplet responses using SNaCK [50] like approach.

Secondly, the non-metric nature of queries creates a 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 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, we made the assumption that the similarity between ideas can be judged even though by definition these ideas (in the form of sketches) are still incomplete and in need of development from an early design concept into a finalized description to build a new device. A more accurate assessment on design items can be given if they are completed (similar to Hollywood pitches in [51]). The similarity assessment will also be affected by the lack of drawing skills of students, which may alter how participants view similarity between ideas. We also asked the participants to assess the similarities at the idea level and not at the feature level. Although, averaging the results of multiple participants provided an estimate of aggregate view of features, the problem is inherently of multiple views for each participant. In future work, we will explore directly optimizing for multiple maps corresponding to each feature using multi-view triplet embeddings [52]. This will allow us to obtain multiple maps for each participant corresponding to different factors considered by them.

7 Conclusion

In this paper, we proposed a method to find idea maps or low dimensional embedding of design ideas using pairwise comparisons. Interpreting these idea maps give insights into what items are considered similar by participants or a group of participants, what attributes are considered important by them in judging similarity and what design items are harder to judge. We show how these idea maps can be used to explain and measure novelty of ideas, where novelty of an idea is measured by how far it is from all other ideas on a map. 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 highlighted interesting properties of how participants chose to differentiate concepts and how to group participants by similarity. They pave the way to use computational methods to reveal what makes ideas novel and allow easy interpretation of results by visualizing ideas on a 2-D map. We compared our results using both a completely automated method and using maps made directly by participants. 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 participant and calculate feature specific novelty.

Acknowledgements

We thank the anonymous reviewers for their efforts in improving the manuscript. 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] Simonton, D. K., 2012. "Taking the us patent office criteria seriously: A quantitative three-criterion creativity definition and its implications". *Creativity research journal,* **24**(2-3), pp. 97–106.

[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] 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.

[9] Seymore, S. B., 2011. "Rethinking novelty in patent law". *Duke Law Journal*, pp. 919–976.

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

[11] 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.

[12] 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.

[13] 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**.

[14] Yu, B., Willis, M., Sun, P., and Wang, J., 2013. "Crowdsourcing participatory evaluation of medical pictograms using amazon mechanical turk". *Journal of medical Internet research,* **15**(6).

[15] Wu, H., Corney, J., and Grant, M., 2015. "An evaluation methodology for crowdsourced design". *Advanced Engineering Informatics,* **29**(4), pp. 775–786.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] Shah, J. J., Smith, S. M., and Vargas-Hernandez, N., 2003. "Metrics for measuring ideation effectiveness". *Design studies,* **24**(2), pp. 111–134.

[22] 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.

[23] 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).

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

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

[26] 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.

[27] 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.

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

[29] 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.
[30] 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.
[31] Torgerson, W. S., 1958. "Theory and methods of scaling.".
[32] 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.
[33] Stewart, N., Brown, G. D., and Chater, N., 2005. "Absolute identification by relative judgment.". *Psychological review,* **112**(4), p. 881.
[34] 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.
[35] 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.
[36] Sankaranarayanan, S., Alavi, A., and Chellappa, R., 2016. "Triplet similarity embedding for face verification". *arXiv preprint arXiv:1602.03418*.
[37] 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.
[38] Haghiri, S., Ghoshdastidar, D., and von Luxburg, U., 2017. "Comparison-based nearest neighbor search". In Artificial Intelligence and Statistics, pp. 851–859.
[39] Ukkonen, A., Derakhshan, B., and Heikinheimo, H., 2015. "Crowdsourced nonparametric density estimation using relative distances". In Third AAAI Conference on Human Computation and Crowdsourcing.
[40] 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.
[41] 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.
[42] Kawakita, J., 1991. "The original kj method". *Tokyo: Kawakita Research Institute*.
[43] 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.
[44] Gower, J. C., 1975. "Generalized procrustes analysis". *Psychometrika,* **40**(1), pp. 33–51.
[45] 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.
[46] Amid, E., Vlassis, N., and Warmuth, M. K., 2016. "Low-dimensional data embedding via robust ranking". *arXiv preprint arXiv:1611.09957*.
[47] Hoffmann, H., 2007. "Kernel pca for novelty detection". *Pattern Recognition,* **40**(3), pp. 863–874.
[48] Ahmed, F., and Fuge, M., 2018. "Ranking ideas for diversity and quality". *Journal of Mechanical Design,* **140**(1), p. 011101.
[49] 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.
[50] Wilber, M., Kwak, I. S., Kriegman, D., and Belongie, S., 2015. "Learning concept embeddings with combined human-machine expertise". In International Conference on Computer Vision (ICCV).
[51] Elsbach, K. D., and Kramer, R. M., 2003. "Assessing creativity in hollywood pitch meetings: Evidence for a dual-process model of creativity judgments". *Academy of Management journal,* **46**(3), pp. 283–301.
[52] Amid, E., and Ukkonen, A., 2015. "Multiview triplet embedding: Learning attributes in multiple maps". In International Conference on Machine Learning, pp. 1472–1480.