Proceedings of IDETC/CIE 2013
ASME 2013 International Design Engineering Technical Conferences &
Computers and Information in Engineering Conference
August 4-7, 2013, Portland, USA

DETC2013-12620

AUTOMATICALLY INFERRING METRICS FOR DESIGN CREATIVITY

Mark Fuge[*]
Berkeley Institute of Design
Dept. of Mechanical Engineering
University of California
Berkeley, CA 94709
Email: mark.fuge@berkeley.edu

Josh Stroud
Berkeley Institute of Design
Dept. of Mechanical Engineering
University of California
Berkeley, CA 94709
Email: jstroud@berkeley.edu

Alice Agogino
Berkeley Institute of Design
Dept. of Mechanical Engineering
University of California
Berkeley, CA 94709
Email: agogino@berkeley.edu

ABSTRACT

Measuring design creativity is crucial to evaluating the effectiveness of idea generation methods. Historically, there has been a divide between easily-computable metrics, which are often based on arbitrary scoring systems, and human judgement metrics, which accurately reflect human opinion but rely on the expensive collection of expert ratings. This research bridges this gap by introducing a probabilistic model that computes a family of repeatable creativity metrics trained on expert data. Focusing on metrics for variety, a combination of submodular functions and logistic regression generalizes existing metrics, accurately recovering several published metrics as special cases and illuminating a space of new metrics for design creativity. When tasked with predicting which of two sets of concepts has greater variety, our model matches two commonly used metrics to 96% accuracy on average. In addition, using submodular functions allows this model to efficiently select the highest variety set of concepts when used in a design synthesis system.

INTRODUCTION

With Design, Creativity, and Innovation increasingly important for competitive advantage, businesses and academics alike are creating various techniques to increase humankind's capability for creativity. As a result, a vast number of books, papers and tools are published every year claiming to increase a person's creativity. In turn, practitioners and researchers need reliable ways of measuring the effectiveness of proposed techniques - in essence, *design creativity metrics*.

However, researchers have yet to reach widespread agreement on appropriate design creativity metrics [1, 2, 3, 4, 5, 6]. Some researchers define the unit of creativity in different ways, including creative design outcomes, design processes, people, and environments [7]. Others draw distinctions between Historical (H) and Psychological (P) creativity [1] (*e.g.*, is an idea novel with respect to all known ideas, or just with respect to an individual?). This paper considers outcome-based metrics (*e.g.*, the novelty of a particular design) judged in a P-creative sense (*i.e.*, from the standpoint of an individual's assessment) [1].

For outcome-based metrics, there have been two primary approaches that past researchers have taken to model creativity: model-based metrics and human judgement-based metrics.

The first approach, model-based metrics, encodes a set of designs into a vector of numbers, which a mathematical formula then evaluates to calculate a score for some aspect of creativity (*e.g.*, variety or novelty). The advantages of model-based metrics are their easy use by both humans and computers in creativity judgement, and their consistency, which encourages reproducible science. The disadvantages are the arbitrary weightings used by some model's scoring systems, and the difficulties in adapting these formalized models to new domains or audiences.

The principle of diminishing marginal utility lies at the core of this class of metrics: the more you have of some design attribute, the less an additional unit is worth. Although this principle is not typically discussed in the context of creativity, it ap-

[*] Address all correspondence to this author.

plies to many aspects creativity. For example, variety, novelty, and unexpectedness all depend on diminishing marginal utility.

The second approach, human-judgement metrics, measures creativity by asking a panel of humans to score concepts using their prior experience. This approach ensures high external validity, but remains expensive to collect and difficult for computational systems to use efficiently.

This paper combines the advantages of these two approaches by proposing a family of easily computable and expressive metrics that can be automatically trained from collections of human judgements. Specifically, it discusses how many existing model-based metrics are based on diminishing marginal utility, and presents a model that ties these metrics together under a general theory. It does so using a special class of functions called *submodular functions* that represent diminishing marginal utility in computationally advantageous ways. Rather than a single metric, our model allows researchers to automatically test and select from an entire family of metrics. This approach essentially finds a specific metric that is best-suited for a given set of human data, also providing a principled method for evaluating among other candidate metrics a researcher might be interested in.

This approach creates strategies that mitigate the two main disadvantages of model-based metrics. By training our model on collections of human judgements, we pair the external validity of human assessment with the computational friendliness and repeatability of model-based metrics. Moreover, by generalizing prior model-based metrics as special cases of diminishing marginal utility, this model allows researchers to adjust existing model-based metrics to better match human assessment.

The main limitation of our approach is that it only models aspects of creativity that exhibit diminishing or constant marginal utility: variety, novelty, unexpectedness are easily modeled, whereas it is not designed to model feasibility, quality, or adherence to requirements. Those aspects of creativity are currently best addressed by other model-based metrics, and combining the two areas is a possible avenue for future work.

Throughout this paper, we use variety metrics as a working example. To validate this approach, experiments demonstrate how the algorithm accurately recovers the existing variety metrics of Shah *et al.* [8] and Verhaegen *et al.* [9] to an average of 97.5% accuracy after 500 binary ratings. The paper also presents results regarding the convergence rate of the algorithm and its robustness under increasing signal-to-noise ratios.

Lastly, our model's use of submodular functions has significant implications for Computational Design Synthesis (CDS) systems that wish to produce creative designs. Notably, a greedy algorithm that selects designs that maximize the proposed submodular function will select the optimally creative set of designs, due to an important connection between creativity and the maximum coverage problem [10]. Through this result, the model provides an efficient means for CDS systems to learn and utilize human creativity when generating new designs.

RELATED WORK

This paper seeks to solve the problem of enabling creative Computational Design Synthesis (CDS) systems, which seek to generate a design, or set of designs, subject to some objective function and constraints [11]. Of particular interest, are systems that produce discrete sets of designs, since in these systems the use of submodular functions can have significant impact. Examples of such CDS systems include Genetic Algorithms (GAs), shape or graph grammar systems [12, 13, 14], agent-based systems [15], and density estimators [16]. In these cases, generating a set of designs typically requires discrete optimization, such as selecting the breeding population in GAs or the production rules in a grammar, to maximize some measure of fitness.

To add creativity to CDS systems, the field needs objective functions that allows model to maximize over aspects of creativity. Some of the model-based metrics mentioned below provide those functions, but do not easily adapt across domains or optimally agree with human judgements. This paper presents a class of convex objective functions that generalizes current metrics and is easy to implement, while also providing a means to adapt to new domains or types of evaluators.

To do so, this paper's contributions build upon two bodies of work: 1) design creativity metrics; and 2) submodular functions.

Model-based Design Creativity Metrics

Outcome metrics that are model-based attempt to mathematically describe the creativity of a set of designs. A critical element in all model-based metrics is some type of formal rubric which allows a person to take a design idea and reliably encode it into a set of numbers summarizing the idea. This encoding process needs to be performed for each concept under consideration, but can often be performed by non-experts provided the rubric is sufficiently well-designed; in contrast, the Judgement-based metrics we review below require expert-level raters for each concept. Although model-based metrics come in many varieties, the most widely used are hierarchical and graph models.

Hierarchical models measure creativity for sets of designs by encoding the set as levels in a tree. An outcome metric, such as variety, is then calculated by measuring various parts of the tree. For example, a popular metric by Shah *et al.* [8] uses a rubric that decomposes concepts into a tree of functions at multiple levels: physical principles, working principles, embodiment, and detail. They then analyze creativity as a combination of four additive sub-metrics: the quantity and variety of the set as a whole, and the quality and novelty of each idea individually [8].

Several researchers have since altered Shah's hierarchical model for various reasons. Nelson *et al.* [17] offer a refined version that fixes several modeling errors. Verhaegen *et al.* [9] combine Shah's metric with a tree entropy penalty, called the Herfindahl index, to encourage "uniformness of distribution" – essentially preferring trees that have even branching. Chakrabarti *et al.* [18] propose to a broader set of functional categories.

Graph-based outcome metrics take a similar approach, but instead of breaking down designs into trees, they compute graph features using attributes like similarity or cluster distances and then combine weighted sums of those features. For example, Maher [19] defines novelty as how far away a new concept is from clusters of previous concepts, where the clusters are created using a prior concept similarity graph.

Human Judgement-based Design Creativity Metrics

Human judgement-based outcome metrics assume that the full extent of what defines creativity cannot be captured in a simple mathematical model. Instead, judgement regarding what is creative is given by a human, usually a domain expert. These metrics typically use a small set of human raters who manually rate designs on a Likert-type scale. The desired outcome metric (*e.g.*, novelty, variety, *etc.*) is then a combination (typically the average) of the human ratings. Metrics that fall under this category include Amabile's Consensual Assessment Technique [20], Carrol *et al.*'s Creativity Support Index [21], and the Creative Product Semantic Scale [22]. Oman *et al.* [23] offer a comprehensive comparison of different metrics, where different methods of evaluation include scale ratings, flow charts, novel models, adjective pairings, and A/B tests.

While human judgement-based metrics have excellent validity (a high score, by definition, is what real humans considered creative), they suffer from two fundamental challenges: reproducibility and expense. Even if it were possible for multiple studies to utilize the same expert raters, differences in knowledge or attitude at time of rating can make evaluators inconsistent with prior ratings. This inconsistentcy makes it difficult to exactly reproduce findings from other papers, even using the same design concepts. Srivathsavai *et al.* [24] found that inter-rater reliability between experts can be low, depending on which aspects of creativity are being evaluated. Collecting expert ratings is also expensive, requiring multiple raters for every concept considered, making judgement of creativity difficult on a large scale.

The model proposed in this paper encompasses a broad range of model-based metrics – it is a *family* of metrics defined by a few free parameters. When these parameters are set to particular values, the model to becomes either a previously published metric or new a metric. Our approach automatically fits these parameters to human judgement data, thereby selecting the particular metric which best matches human judgements. This requires a small, one-time collection of expert-level human evaluation data, but then pays off with a reproducible model-based metric with high external validity that can be used by non-experts provided they use a rubric that can encode designs into a set of numbers. In essence, our model is a middle-ground between model-based and human judgement-based metrics.

Submodular Functions

We use submodular functions as a fundamental tool to model and use creativity in an efficient way. For example, say we add an item x to a set of items A; a function is submodular if we get a greater increase in value from adding x to A, than adding x' to the set $\{A \cup x\}$. In short, the more items we add to A, the less each additional item is worth. A common example of a submodular function is the logarithm (for each positive δx we move, δy decreases). This definition is where submodular functions gain their usefulness: it is identical to the principle of diminishing marginal utility. The formal definition is that submodular functions are set-based functions where, for a function ρ and two sets $A, B \in \Omega$: $\rho(A) + \rho(B) \geq \rho(A \cup B) + \rho(A \cap B)$. This definition is similar to the behavior of the logarithm as described above, except that A and B are sets, rather than a continuous variable x.

Recently, machine learning researchers have adapted submodular functions to solve large-scale problems, particularly in developing algorithms that recommend an optimally diverse set of relevant webpages during a search. These opportunities lead to formally defining the idea of "coverage" for a set of documents as the extent to which a set of items covers all possible elements. Finding maximum coverage is called the Maximum Coverage Problem, and has been proven to be NP-Hard.

Khalid *et al.* [25] demonstrated how submodular functions could produce diverse webpage results, since lower bounds on the performance of submodular functions [10, 26] provide the optimal approximation for solving the Maximum Coverage Problem. Since that time, others have built upon the use of submodular functions for diverse retrieval, notably the work by Ahmed *et al.* [27], upon which our model is based.

By demonstrating the connection between creativity, diminishing marginal utility, and submodular functions, this paper allows the design community to make use of advances in other fields to develop better creativity metrics and CDS systems.

CREATIVITY MODEL

This paper's core insights lie in the following connections, resulting in the approach shown in Fig. 1:

1. Many common elements of creativity, such as variety or novelty, are naturally expressed via the principle of diminishing marginal utility.
2. Diminishing marginal utility can be expressed in a computationally advantageous way via submodular functions.
3. Submodular functions can easily utilize many of the design representations used in current creativity metrics.
4. Given a set of designs, human experts have a hard time agreeing on real numbered values for its creativity, but easily make binary "greater than" or "less than" judgements.

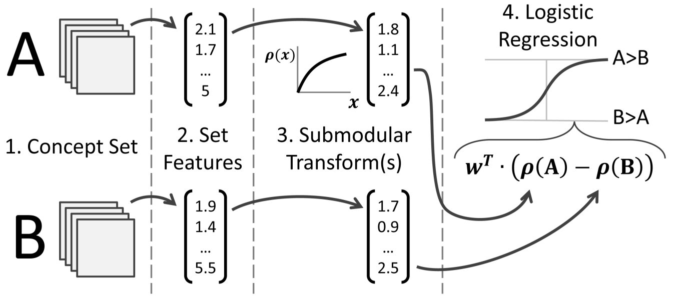

FIGURE 1. The overall approach 1) takes in two set of design concepts (A and B), 2) encodes each set into a vector of features, 3) transforms those features through submodular functions ($\rho(A)$, $\rho(B)$), and 4) determines which set has greater variety using a weighted ($\mathbf{w}$) difference ($\rho(A) - \rho(B)$) between the submodular features of the two sets. A logistic regression optimizes the weight of each submodular feature ($\mathbf{w}$) so that the model closely matches expert-rated comparison data.

Connecting Creativity, Diminishing Marginal Utility, and Submodular Functions

To see how creativity, diminishing marginal utility, and submodular functions are related, we return to the example of estimating variety. We use a variant of the example presented in Shah *et al.* [8]: suppose we have a set of student-generated designs whose purpose is to move an object from point A to point B. We want to select the two designs from that set which have the most variety. For simplicity, assume that we have just three designs: 1) a small cart that propels itself forward using a balloon filled with air; 2) a similar cart, but propelled using a balloon filled with water; and 3) a small catapult.

Say that we choose the first cart as one of your two final choices: which of the other two designs do we pick? Since we already have a balloon-propelled cart design, we do not get much value from picking the second cart. This additional value is our marginal utility, which diminishes because the second cart design is not as valuable as the first (even if they are equally good designs). On the other hand, selecting the catapult to go along with the first cart would give us higher marginal utility, since a catapault is a completely new way of transporting the object and thus has higher variety.

Various metrics try to address this notion of diminishing marginal utility. Hierarchical metrics proposed by Shah *et al.* [8] and subsequent work [18, 17] represent this principle by assigning a higher reward for solutions at higher functional levels. Verhaegen *et al.* [9] take those metrics a step further by accounting for the entropy of the concept distribution, which is similar in purpose to diminishing marginal utility. Maher [19] models it as a reward for greater aggregate distance from existing cluster centers. Whether a discrete or continuous space, the idea remains the same: if a new idea is similar to what you already have, it is less valuable – that is, it has a diminished marginal utility.

As we showed above, submodular functions closely model diminishing marginal utility, which we can use to measure creativity. To operationalize this new knowledge, we need to address the following questions: 1) how are aspects of creativity expressed as submodular functions, 2) how do we represent designs for use in submodular functions, and 3) how do we use those functions to emulate human judgments?

Modeling Creativity with Submodular Functions

As with most published creativity metrics, we use a linear model where the outcome metric is modeled as a vector of weights multiplied by a vector of features. Returning to variety as the example: variety$(A) = \mathbf{w}^T \cdot \mathbf{d}(A)$, where A is the set of designs, $\mathbf{d}$ is a vector of numbers summarizing the features of A, and $\mathbf{w}$ is a vector of weights for each feature. In prior metrics the feature weights ($\mathbf{w}$) are typically set to some constant value (*e.g.*, Shah *et al.* [8] choose $\mathbf{w} = [10, 6, 3, 1]$).

This is where our work departs from prior work. We use submodular functions ($\rho(x)$) to transform the design features such that they obey specific forms of diminishing marginal utility: variety$(A) = \mathbf{w}^T \cdot \rho(\mathbf{d}(A))$. We apply a variant of the model of Ahmed *et al.* [27] for the purposes of modeling design creativity.

Formally, the submodular score for a set A is given by

$$f(A) = \boldsymbol{\gamma}^T \cdot \mathbf{d}(A) + \boldsymbol{\beta}^T \cdot \rho(\mathbf{d}(A)) \quad (1)$$

where the weight vector $\mathbf{w}$ has been broken into two parts: $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, representing the modular and submodular contributions to variety, respectively. This arrangement allows the algorithm to determine how whether a feature obeys diminishing marginal utility. Either $\boldsymbol{\gamma}$ or $\boldsymbol{\beta}$ can be set to zero to use only the submodular or modular parts, respectively. This paper assumes that variety behaves fully sub-modularly, so that we set $\boldsymbol{\gamma} = 0$, resulting in $f(A) = \boldsymbol{\beta}^T \cdot \rho(\mathbf{d}(A))$. In this paper $\rho(\mathbf{d}(A))$ is a vector where ρ has been applied to each element in the vector $\mathbf{d}(A)$.

While any submodular function can be used for ρ, some useful options given by Ahmed *et al.* [27] include:

Set Cover: $\rho(x) = 1$ if $x > 0$; 0 if $x = 0$
Probabilistic Cover: $\rho(x) = 1 - e^{-\theta x}$ for $\theta > 0$
Logarithmic Cover: $\rho(x) = \log(\theta x + 1)$ for $\theta > 0$

This paper demonstrates in the experiment section below that the metrics of Shah *et al.* [8] are a special case of this paper's model where $\rho =$ set cover, while the metric of Verhaegen *et al.* [9] is well approximated by $\rho =$ probabilistic cover.

Encoding Design Concepts

Now that we have some candidate submodular functions, the next step is to define $\mathbf{d}(A)$, *i.e.*, how a specific set of designs (A) becomes a vector of numbers that can be used by the submodular

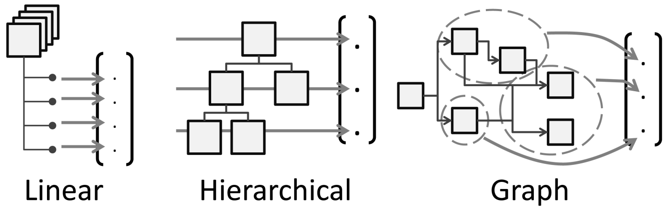

FIGURE 2. This model can encode more than just hierarchical creativity metrics; it can handle graph-based creativity features, over which linear and hierarchical features are a subset.

function. We refer to this process as *encoding* the design concepts, and it is required for any model-based metric. Typically this is done using a rubric that describes how a human evaluator should take a design and summarize it into a list of real numbers. For example, in Shah *et al.* [8, Tab. 6-7] a set of designs is encoded as a functional tree decomposition, where each of four levels is summarized by the number of bifurcations in a particular level of the tree; counting these branches provides you with four numbers that describe the set.

Various authors have proposed different encodings, but the end result is the same: designs become a vector of real numbers ($\mathbf{d}(A)$) (we call these *features*) that get used in a linear model. Our model is agnostic to the choice of rubric or encoding, leaving the researcher free to try whatever rubric they believe accurately captures the aspects of design they are interested in. For human-generated concepts this encoding is performed manually for each concept, while computationally generated designs are typically encoded automatically through a fixed algorithm. While needing to manually encode each concept is disadvantageous, it is a limitation shared across all model-based metrics.

Given a particular encoding, this paper's model determines how each of the encoded features impacts creativity. As Fig. 2 demonstrates, this approach works for linear design encoding, such as those commonly used in consumer preference models; hierarchical encodings, such as Shah *et al.* [8] and others [18, 17, 9]; and graph-based encodings, such as the cluster model of Maher [19] or the Function-Behavior-Structure (FBS) model [28].

By choosing the appropriate encoding, our model can measure the creativity of any aspect of design, and replace many existing metrics. Our model can also be used to compare different encodings and determine which is best for a given problem – a strategy we revisit later in the discussion section.

Model Inference

Given the above model and a particular encoding, the next task is to estimate the weights **w** and any hyperparameters (*e.g.*, θ) using a dataset of human given ratings. Given perfect human raters, we could calculate these properties by asking human judges: "on a scale from 0-10, how much variety does this concept set have?" Unfortunately, this task is unfeasible in practice, since every judge has a different definition of variety, making simple numerical answers difficult to compare across judges.

Instead, we can ask a human evaluator to compare two sets: "Given a set of concepts A and another set B, which set has greater variety?" The result is a binary "$A > B$" or "$A < B$" answer which is easily comparable across raters and is more accurate than absolute value scaled scores [29]. Although this method is still prone to differences in opinion and background (as are all creativity metrics that depend on human evaluation), it reduces differences in absolute measurement between individuals. Possible alternatives to binary rating include ordinal ranking of more than two sets, as well as ordinal categories, *e.g.*, "high", "medium", and "low" variety. These, among other options, can easily be translated to binary greater than/less than judgements that our approach can use when greater fidelity data is available.

Given a dataset of binary judgements (*i.e.*, is $A > B$ or not) between various pairs of sets, standard logistic regression can determine the optimal weights **w** that best match the judgements given by human experts. Formally, the likelihood function for predicting whether a human would rate a set $A > B$ is given by:

$$\mathbf{P}(A > B|A,B) = \left[1 + e^{-(f(A)-f(B))}\right]^{-1} \tag{2}$$

where $f(A)$ and $f(B)$ are given by Eqn. 1. Using the entire dataset, maximum likelihood estimation on the above likelihood function determines the optimal weights. The value of the hyperparameters, if needed, can be determined either through grid search or through stochastic gradient descent for certain forms of the submodular function. Maintaining the model is also simple: if new data is collected after initial training, the model can be easily updated using any sequential gradient descent algorithm, since maximum likelihood for logistic regression is an unconstrained convex optimization problem.

The end result after fitting the model to the human judgement data is an optimized vector of weights **w**, and, optionally, any hyper-parameters (θ). These quantities can then be used in Eqn. 1 to produce a numerical score ($f(A)$) for a set of new designs. Alternately, you could also calculate the predicted human judgement between two sets of designs ($P(A > B|A,B)$) by using the optimized parameters in both Eqns. 1&2.

EXPERIMENTAL RESULTS

Validate our approach, we trained our model to predict which of two randomly generated concept sets had greater concept variety, and recorded whether the model made correct predictions on new data. To do this, we generated synthetic human judgement data by using two different existing model-based metrics, Shah *et al.* and Verhaegen *et al.* to simulate users. By using these metrics to simulate human judgements, we could assess how closely the model uncovered the true judgements, while also making it easy for others to reproduce these results. Our full experiment code is available, for those who to wish to replicate or extend our results: `www.markfuge.com/research/creativity.html`. Work is currently under way to apply this model to collections of human ratings.

For our synthetic dataset, we chose Shah's metric because of its broad adoption and because it is useful example of how hierarchical metrics are encoded as a linear model. We chose Verhaegen's metric since it attempts to account for "uniformness of distribution" within the hierarchical branches using the Herfindahl index. This is similar in spirit to modeling diminishing marginal utility, and makes that metric a natural candidate with which to assess the utility provided by different submodular functions.

To create the synthetic dataset, we leveraged the fact that both Shah's and Verhaegen's metrics involve hierarchical metrics defined on trees of depth $D = 4$, where random concept set generation amounts to randomly generating functional trees for sets of M concepts. These randomly generated function trees were the objects used when creating the $A > B$ binary judgements. In practice, actual human ratings would be used in place of the simulated data, with the researcher free to determine what encoding they are interested in. Function trees are used here only to be consistent with the encodings used by Shah's and Verhaegen's metrics.

The results in this section were generated using the following experimental procedure:

1. Select a variety metric to simulate human judgements (*i.e.*, Shah or Verhaegen).
2. Randomly generate multiple sets of $M = 10$ concepts and calculate their variety with respect to the chosen metric. These values are the ground truth variety scores $(V(X))$.
3. Transform the feature vector of each set of concepts using a submodular function (set cover for Shah, probabilistic cover for Verhaegen) - This transformation is the function $\rho(X)$.
4. Use Eqn. 1 to determine the submodular difference vector between two sets of concepts - $\mathbf{x}_i = \{\rho(A) - \rho(B)\}$ in the case of fully submodular features. These difference vectors become the input features for the logistic regression.
5. Recall the ground truths for each set (A & B). The equation $y_i = \text{sign}(V(A) + \varepsilon_A > V(B) + \varepsilon_B)$ decides whether the variety of A is greater than B, where $\varepsilon = N(0, \sigma^2)$. This decision becomes the classification label for the logistic regression.
6. Steps 4 & 5 are repeated for as many training samples as desired ("# of A/B Comparisons" in Figs. 3-5).
7. Using the difference vectors from step 4 and corresponding classification labels from step 5, use logistic regression to learn the optimal weights ($\mathbf{w}$).
8. Evaluate the model on unseen test data via 30 randomized cross-validation trials, comparing predicted decisions with ground truth labels to determine prediction accuracy.

Metric	n=100	200	300	400	500
Shah	95.2	96.9	97.6	98.1	98.6
Verhaegen	91.9	94.4	95.5	96.0	96.3

TABLE 1. Prediction accuracy across metrics. Randomly guessing achieves a baseline score of 50%.

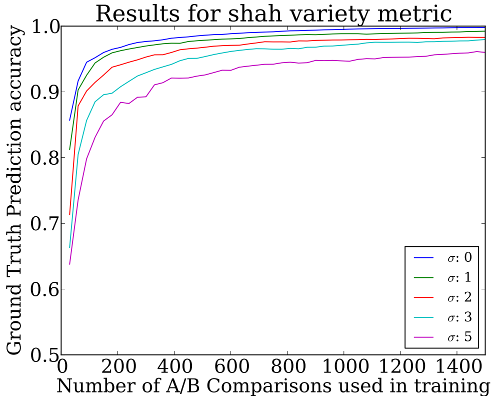

FIGURE 3. With no error (top-most curve), our approach recovers the Shah metric to within 95% by 100 ratings. When more random error is added, the algorithm's convergence is slower. Randomly guessing results in a prediction accuracy of 0.5 (50%).

Figures 3 and 4 demonstrate the convergence and robustness results of the model under Shah *et al.*'s and Verhaegen *et al.*'s metrics, respectively. As Table 1 shows, in both cases the algorithm converges to above 90% accuracy within the first 100 ratings, and to above 95% accuracy within the first 300 ratings. Changing M demonstrated no meaningful change in any results, which matches expectations. At $N = 500$ binary ratings in the no-noise condition, for $D = \{4, 10, 25, 50\}$ the resulting accuracies were $\{98.9, 97.5, 95.0, 93.2\}$ respectively, again matching expectations; the performance curves look similar to those in Fig. 3, but were omitted for space. In the presence of noise, the convergence rate is slower but the model is able to recover the underlying metric to high accuracy, given sufficient training samples.

Figure 5 demonstrates how the choice of submodular cover type affects the recovery accuracy of the model: for Shah's metric, using set cover makes the model equivalent, and thus it captures the metric with complete accuracy. Using probabilistic cover reduces the accuracy, since Shah's metric does not encode diminishing marginal utility within each level of the tree. Under Verhaegen's metric, using set cover does not capture the model as accurately as using probabilistic cover, since their metric does

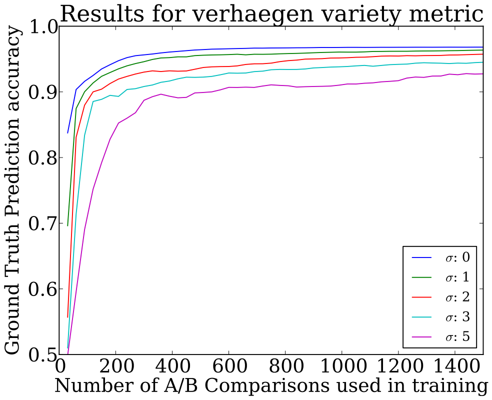

FIGURE 4. Our model recovers Verhaegen's metric to within 95% by 300 ratings. The average scores are lower, since, unlike Shah's metric, our model class does not perfectly contain Verhaegen's metric.

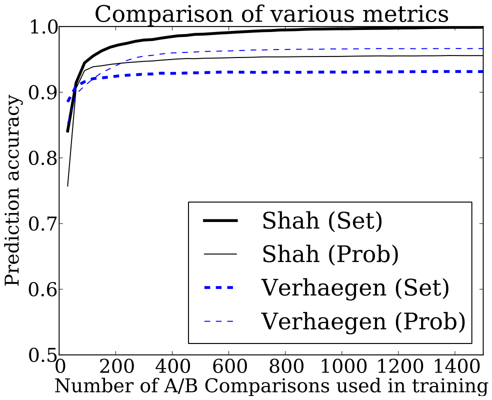

FIGURE 5. Comparison of different submodular set cover types. Shah's metric performs better under set cover, where Verhaegen's performs better under probabilistic cover.

attempt to encode diminishing marginal utility within each level of the tree. Both of these results confirm expectations. Logarithmic cover and probabilistic cover achieve similar results in both cases, so we present only probabilistic cover to improve figure clarity.

DISCUSSION

The above results raise the following questions for discussion:

1. To what extent does this model extend to aspects of creativity other than variety?
2. How would these results fare under actual human evaluation, instead of simulated sources?
3. What do these results mean for other work in Design Creativity measurement?
4. How does this model affect Computational Design Synthesis systems?

Extensions to Other Aspects of Creativity

This paper provides two possible avenues for extending metrics: 1) testing other structures and domains for variety, and 2) modeling other aspects of creativity, such as novelty and usefulness. In each case, the resulting mathematical model and inference procedures remain unchanged: the only change is how the model encodes design concepts. This leads to a general procedure for experimenting with various encodings and datasets.

The experiments in this paper focused on hierarchical metrics that are commonly used to analyze variety in engineering design. However, this model extends to any encoding that can be expressed as a linear set of features. This extension opens up future work in formulations of variety on existing domains, as well as transferring existing metrics to different domains.

For example, a new graph-based metric could be applied to studying sets of FBS models or patent networks to determine which features accurately predict variety. Likewise, by training on a different set of experts, a hierarchically structured metric similar to Shah *et al.*'s could be adapted to describe functions in an organizational or service context.

The presented model can also be adapted to describe any creativity metric that utilizes a form of diminishing marginal utility. For example, novelty (*e.g.*, [8, 18, 19]) or unexpectedness (*e.g.*, [19]) are recreated by altering which training sets are used.

Novelty would be the marginal utility between the current set A, and a new set $B = \{x \cup A\}$. Judges could rate two sets to determine what aspects affect novelty. Unexpectedness can be formulated similarly; unlike novelty, however, unexpectedness only considers the most recently seen designs. The unexpectedness model essentially "forgets" old designs over time, and becomes "surprised" if something breaks a chain of similar designs.

As mentioned earlier, a limitation of our model is that it is not designed to model aspects of creativity that do not exhibit constant or decreasing marginal utility. In cases where including those aspects is desired, we recommend using our model in concert with other good model-based metrics that cover those aspects. We have made our model code freely available to provide a platform for future work in this area.

The Utility of Human Evaluation

The proposed model requires the collection of human ratings, raising a natural concern: If the ratings are noisy, or even contradictory, will this model be of any use? What if humans are consistently poor judges of a certain aspect of creativity?

Fig. 3 and 4 provide an answer to this question: the model handles noise gracefully, even if the raters' assessments differ by a large amount. Increased noise translates to increased convergence time, but even under high levels of noise ($N(0,5^2)$ for a 10 point variety metric), the model can combine multiple ratings and uncover the underlying variety score to within 95% accuracy.

These experimental results assume that experts' ratings are normally distributed around a "true" variety score. This assumption is not quite true, but reasonably approximates reality and allows us to offer these initial robustness results.

The proposed model naturally accounts for differences between individual raters or groups of raters. By extending the score function (Eqn. 1) with a set of user or group-specific bias terms, this model can automatically learn these differences given additional training data. This approach is commonly used to capture of possible bias terms in linear models (*e.g.*, [30]).

If human judgements contradict each other, or if expert judgements are consistently wrong about a set of metrics, then the proposed model will mirror that behavior. However, this case can be easily checked since the model can provide confidence estimates for its accuracy (a non-trivial task for a human rater).

A relevant issue is the selection of an appropriate population of human judges. Do we need domain experts and professional designers, or can we settle for non-experts? This question is best answered through appropriate controlled studies, such as the one conducted by Kudrowitz and Wallace [31], who demonstrated that Mechanical Turk raters showed strong correlation with novelty but poor correlation with feasibility.

Lastly, the convergence behavior suggests the number of ratings required to train the model. Both Shah and Verhaegen's variations reached or exceed 90% accuracy within 500 samples, even under extremely noisy conditions. This level of convergence could be achieved with 10 raters, who each supply 50 ratings. Up to an order of magnitude more data could easily be collected in practice, implying that our approach is feasible. Once the model is trained, it requires no additional expert data achieve results, unlike traditional methods. When a researcher wishes to estimate additional creative factors, our results on increasing D demonstrate that the amount of data needed increases, but not prohibitively so.

Impact on Design Creativity Measurement

The generalizability of the proposed approach opens up many new questions and future work opportunities for those working in design creativity measurement:

To what extent does diminishing marginal utility occur in aspects of creativity? Figure 5 demonstrates how to identify the presence of diminishing marginal utility: the group of users simulated by Verhaegen's metric were more accurately modeled by introducing diminishing marginal utility across the function tree branches, while those simulated by Shah's metric were not. This suggests a method for systematically investigating which attributes of designs obey diminishing marginal utility – evaluate different models with different types of submodular functions to hypothesize possible models for creative behavior.

Likewise, researchers can try different encodings (*e.g.*, linear, hierarchical, or graph structured) to determine which model best matches the creativity judgements provided by experts. By using the same set of human judgements, new and published metrics can be assessed for how closely they match reality. Our approach creates a feedback loop for hypothesis-driven creativity research that enables the research community to to systematically select and develop more accurate creativity metrics.

Are some features more important to creativity than others? In order to use the proposed approach to evaluate how important different design attributes are for creativity, we can do one of two things: 1) compare a large number of different design encodings, determine which one best fits human data, and then inspect that model's weights ($\mathbf{w}$) to determine importance, or 2) create a design encoding with as many design features as possible, and then train the model using L1 regularization in Eqn. 2 to encourage unimportant weights to be driven to zero. In addition, new computational algorithms can be derived to identify important features not yet known: algorithms that cluster ideas according to diminishing marginal utility could be given to domain experts to uncover patterns in human evaluation.

Do different domains, experience levels, or backgrounds judge the same creativity metric differently? By using a particular design encoding and training the model on different groups of people, future work could formalize differences in opinion regarding the same metric. For example, given a Shah-like variety metric, would architects and engineers view the importance of physical function differently? Comparing the learned weights of the metric for each group could provide an answer.

Impact on Computational Design Synthesis Systems

The use of submodular functions has several advantages for CDS systems that wish to optimize over creativity:

The objective function is convex in the input features. The convexity of Eqn. 1 has obvious advantages when optimizing over a continuous design space.

The objective function in Eqn. 1 can be easily incorporated in multi-objective optimization. After training the creativity model using logistic regression (Eqn. 2), the submodular function and weights (Eqn. 1) can be reused separately to provide a variety score. This score can be used inside of a multi-objective optimization loop to balance creativity with other performance objectives.

Finding the most creative set of designs is an NP-Hard problem, but our model offers the best possible approximation guarantee. Selecting the highest variety set of concepts (and by above extension, highest novelty or most unexpected) is equivalent to the Maximum Coverage Problem. This means that CDS algorithms will not be able to efficiently select the most creative set in polynomial time, and any polynomial time algorithm can only approximate the solution to $\approx 63.2\%$ or less of optimum [10]. This seems to paint a daunting picture for the future success of creative CDS systems.

Thankfully, the use of submodular functions provides relief: a greedy algorithm that sequentially selects the designs that maximize Eqn. 1 is guaranteed to approximate the Maximum Coverage solution to at least $1-\frac{1}{e} \approx 63.2\%$ of optimum [10]. This essentially matches the upper bound on the approximate solution of the Maximum Coverage Problem, meaning that our approach provides the best possible approximation you can hope for when attempting to optimize the creativity of design sets. For proof regarding this optimality or details about the greedy selection algorithms that achieve that optimality, we direct readers to the following papers: [10, 32]. This approximation provides significant cost savings when the set of possible designs is large, such as in CDS systems that automatically generate designs [11].

CONCLUSIONS

The strength of this paper lies in drawing an important theoretical connection between certain aspects of creativity, such as novelty and variety, and the principle of diminishing marginal utility. By utilizing submodular functions to express diminishing marginal utility, this paper described a creativity model that ties together many existing metrics under a common framework.

Our model generalizes different configurations of creativity metrics, such as linear, hierarchical, or graph based metrics. The model can also adapt to human evaluators from different backgrounds. It does so by requiring only simple A/B comparisons between sets of concepts, simplifying data collection with a rating task easily processed by human judges.

As validation, this paper demonstrated how the proposed model can reliably predict judgements produced by simulating two published creativity metrics. Using the variety metrics of Shah *et al.* [8] and Verhaegen *et al.* [9] to simulate judgement data, the model predicted future judgements with 100% and 96.4% accuracy, respectively. Under increasingly noisy input conditions, the model is still able to recover the judgements accurately, at the cost of some convergence speed.

The use of submodular functions to model diminishing marginal utility carries with it several advantages: 1) the model parameters can be interpreted easily, 2) the likelihood and objective function are convex allowing for efficient optimization, and 3) Computational Design Synthesis systems can use the model to perform optimal set selection in an efficient way.

These strengths come with a major limitation: there are several aspects of creativity that do not have the diminishing marginal utility property, such as feasibility or quality. While our approach of sub-modular functions cannot be used to capture these aspects of creativity, the proposed model can be used in concert with other model-based metrics that address those aspects. Future work could apply our data-driven approach to measurement of creativity metrics across a broad spectrum of areas.

Rather than claiming to provide a universal metric for creativity, this work instead presents a family of metrics that can act as a catalyst with which design creativity researchers can ask new questions:

1. How does human evaluation of a particular creativity metric vary across different conditions (disciplines, countries, professional experience, *etc.*)?
2. What are the important elements that determine creativity? What kinds of model structure appropriate? To what extent can we discover those structures given human rating data?
3. How can Computational Design Synthesis systems utilize models of creativity to generate creative designs? What are the most efficient ways for CDS systems to query human evaluators to best emulate creative design?

By combining the reproducibility of mathematical models with the credibility of human judgements, this paper allows designers and researchers access to more robust, adaptable, and externally valid ways of quantifying creativity.

ACKNOWLEDGEMENTS

The authors are grateful for the partial support from NSF grant IIS-0856098 and from the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program. They also thank the reviewers for providing valuable feedback on the manuscript.

REFERENCES

[1] Boden, M., 2009. “Computer Models of Creativity”. *AI Magazine,* **30**(3), pp. 23–34.

[2] Brown, D. C., 2011. “The Curse of Creativity”. In *Design Computing and Cognition ’10*, J. S. Gero, ed. Springer Netherlands, Dordrecht, ch. 9, pp. 157–170.

[3] Christiaans, H., and Venselaar, K., 2005. “Creativity in Design Engineering and the Role of Knowledge: Modelling the Expert”. *International Journal of Technology and Design Education,* **15**(3), Jan., pp. 217–236.

[4] Gero, J., 2000. “Computational Models of Innovative and Creative Design Processes”. *Technological Forecasting and Social Change,* **64**(2-3), June, pp. 183–196.

[5] Burkhardt, J.-M., and Lubart, T., 2010. “Creativity in the Age of Emerging Technology: Some Issues and Perspec-

tives in 2010". *Creativity and Innovation Management,* **19**(2), pp. 160–166.

[6] Puccio, G. J., Cabra, J. F., Fox, J. M., and Cahen, H., 2010. "Creativity on demand: Historical approaches and future trends". *AI EDAM,* **24**(Special Issue 02), pp. 153–159.

[7] Saunders, R., and Gero, J. S., 2002. "How to Study Artificial Creativity". In Proceedings of the 4th conference on Creativity and cognition, ACM Press, pp. 80–87.

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

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

[10] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L., 1978. "An analysis of approximations for maximizing submodular set functions - I". *Mathematical Programming,* **14**(1), Dec., pp. 265–294.

[11] Chakrabarti, A., Shea, K., Stone, R., Cagan, J., Campbell, M., Hernandez, N. V., and Wood, K. L., 2011. "Computer-Based design synthesis research: An overview". *Journal of Computing and Information Science in Engineering,* **11**(2), pp. 021003+.

[12] McCormack, J. P., Cagan, J., and Vogel, C. M., 2004. "Speaking the buick language: capturing, understanding, and exploring brand identity with shape grammars". *Design Studies,* **25**(1), Jan., pp. 1–29.

[13] Campbell, M. I., Rai, R., and Kurtoglu, T., 2012. "A stochastic Tree-Search algorithm for generative grammars". *Journal of Computing and Information Science in Engineering,* **12**(3), pp. 031006+.

[14] Talton, J., Lou, Y., Lesser, S., Duke, J., Měch, R., and Koltun, V., 2011. "Metropolis procedural modeling". *ACM Trans. Graphics,* **30**(2), April.

[15] Campbell, M. I., Cagan, J., and Kotovsky, K., 1999. "A-Design: An Agent-Based approach to conceptual design in a dynamic environment". pp. 172–192.

[16] Talton, J. O., Gibson, D., Yang, L., Hanrahan, P., and Koltun, V., 2009. "Exploratory modeling with collaborative design spaces". In Proceedings of the 2nd Annual ACM SIGGRAPH Conference and Exhibition in Asia, ACM Press.

[17] Nelson, B. A., Wilson, J. O., Rosen, D., and Yen, J., 2009. "Refined metrics for measuring ideation effectiveness". *Design Studies,* **30**(6), Nov., pp. 737–743.

[18] Sarkar, P., and Chakrabarti, A., 2011. "Assessing design creativity". *Design Studies*, Mar.

[19] Maher, M. L., 2010. "Evaluating creativity in humans, computers, and collectively intelligent systems". In Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design, DESIRE '10, Desire Network, pp. 22–28.

[20] Amabile, T. M., 1982. "Social psychology of creativity: A consensual assessment technique". *Journal of Personality and Social Psychology,* **43**, pp. 997–1013.

[21] Carroll, E. A., Latulipe, C., Fung, R., and Terry, M., 2009. "Creativity factor evaluation: towards a standardized survey metric for creativity support". In Proceeding of the seventh ACM conference on Creativity and cognition, C&C '09, ACM, pp. 127–136.

[22] O'Quin, K., and Besemer, S. P., 1989. "The development, reliability, and validity of the revised creative product semantic scale". *Creativity Research Journal,* **2**(4), pp. 267–278.

[23] Oman, S., Tumer, I., 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**, pp. 65–92.

[24] Srivathsavai, R., Genco, N., Holtta-Otto, K., and Seepersad, C. C., 2010. "Study of existing metrics used in measurement of ideation effectiveness". In Volume 5: 22nd International Conference on Design Theory and Methodology; Special Conference on Mechanical Vibration and Noise, ASME, pp. 355–366.

[25] El-Arini, K., Veda, G., Shahaf, D., and Guestrin, C., 2009. "Turning down the noise in the blogosphere". In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD).

[26] Krause, A., and Guestrin, C., 2011. "Submodularity and its applications in optimized information gathering". *ACM Trans. Intell. Syst. Technol.,* **2**(4), July.

[27] Ahmed, A., Teo, C. H., Vishwanathan, S. V. N., and Smola, A., 2012. "Fair and balanced: learning to present news stories". In Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12, ACM, pp. 333–342.

[28] Gero, J. S., 1990. "Design prototypes: a knowledge representation schema for design". *AI Mag.,* **11**(4), Oct., pp. 26–36.

[29] Carterette, B., Bennett, P., Chickering, D., and Dumais, S., 2008. "Here or there". In *Advances in Information Retrieval*, C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. White, eds., Vol. 4956 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 16–27.

[30] Koren, Y., 2009. "The BellKor Solution to the Netflix Grand Prize".

[31] Kudrowitz, B. M., and Wallace, D., 2012. "Assessing the quality of ideas from prolific, early-stage product ideation". *Journal of Engineering Design*, Apr., pp. 1–20.

[32] Krause, A., Leskovec, J., Guestrin, C., VanBriesen, J., and Faloutsos, C., 2008. "Efficient sensor placement optimization for securing large water distribution networks". *Journal of Water Resources Planning and Management,* **134**(6), November, pp. 516–526.