Proceedings of the ASME 2017 12th International Manufacturing Science and Engineering Conference
MSEC2017
June 4-8, 2017, Los Angeles, CA, USA

MSEC2017-2790

DEVELOPING A CAPABILITY-BASED SIMILARITY METRIC FOR MANUFACTURING PROCESSES

Kevin Li
Department of Mechanical Engineering
University of Maryland
College Park, Maryland 20742

William Z. Bernstein*
Systems Integration Division
National Institute of Standards and Technology
Gaithersburg, Maryland 20899

KEYWORDS
Unit manufacturing process; process capability;
similarity metric; supplier discovery; database exploration

ABSTRACT
Manufacturing taxonomies and accompanying metadata of manufacturing processes have been catalogued in both reference books and databases on-line. However, such information remains in a form that is uninformative to the various stages of the product life cycle, including the design phase and manufacturing-related activities. This challenge lies in the varying nature in how the data is captured and represented. In this paper, we explore measures for comparing manufacturing data with the goal of developing a capability-based similarity metric for manufacturing processes. To judge the effectiveness of these metrics, we apply permutations of them to 26 manufacturing process models, such as blow molding, die casting, and milling, that were created based on the ASTM E3012-16 standard. Furthermore, we provide directions towards the development of an aggregate similarity metric considering multiple capability features. In the future, this work will contribute to a broad vision of a manufacturing process model repository by helping ease decision-making for engineering design and planning.

1 INTRODUCTION
Digital manufacturing has fundamentally changed the way in which organizations design, build, and assess products. The wealth of manufacturing-related data has been exploited in many ways, to date. According to a report by McKinsey Global Institute, "manufacturing stores more data than any other sector – close to 2 exabytes of new data stored in 2010" [1]. However, it is widely accepted that until now, the manufacturing world is far from meeting its true potential in the digital age [2]. This wealth of data unfortunately lacks sufficient context, which is negatively affecting its realized value. All in all, the manufacturing environment has become data rich yet information poor.

One challenge for (semi-)automating manufacturing decision making is properly representing tacit knowledge of manufacturing experts. Currently, manufacturing databases house general information and do not provide recommendations for complex decision scenarios. An example of such a decision scenario is finding a manufacturing supplier that meets specifications for a new design, or in short, supplier discovery. In this case, the decision maker must balance the capabilities of a single supplier across a variety of dimensions, e.g. produced shape, achievable tolerance and surface roughness. Narrowing down a particular manufacturing process based on capabilities is often left to the discretion of the human and their experience. Studies have shown that the coordination between machine-driven algorithms and human-based intuition improves decision making processes [3]. Similarly, one goal of this work is to enhance process discovery through human-machine coordination.

In this paper, we explore various techniques for defining the similarity of two manufacturing processes in the context of supplier discovery. The eventual goal of this research is to develop a general metric that enables faster and easier comparison of manufacturing process capabilities. We can envision that such a metric would complement existing efforts in ontology and linked

*Address all correspondence wzb@nist.gov

data development to enhance design and manufacturing engineers' toolboxes to make better decisions. This paper proposes several metrics for different manufacturing attributes depending on the nature of the data. Furthermore, we present initial work towards a unified metric that takes into account multiple manufacturing capabilities in a weighted scheme.

The rest of the paper is organized as follows. Section 2 discusses relevant work from the perspective of standards, databases, and similarity metrics. Section 3 describes the script that was used to parse the data and the metrics that were adapted and developed to compare the processes. Section 4 presents results after computing the various metrics using a test dataset along with its interpretation. Section 5 addresses limitations of our approach, primarily focused on data-related issues. Section 6 looks at future work for the project and how it can be implemented into a larger initiative to facilitate the manufacturing design process. The vision of this work is to include the metric and the related algorithms within a structured database to enable better query mechanisms for human decision makers.

2 BACKGROUND & MOTIVATION

This section reviews related work with respect to (1) standardizing information models describing manufacturing systems, (2) storing information related to capabilities of manufacturing processes, and (3) developing similarity metrics for such information models. Here, we motivate our work for constructing an automated measure of similarity of manufacturing processes.

2.1 Standardizing manufacturing information models

Primary efforts in standardizing information models that formally characterize manufacturing processes include IS0 20140 [4] and ASTM E3012-16 [5], both of which focus on environmental considerations of manufacturing processes. Both standards address the need for improving environmental models to populate life cycle inventories (LCI). In fact, ISO 20140 specifically states that its proposed reference model directly aligns with the EcoSpold *de facto* standard that has been widely adopted for storing LCI and life cycle assessment (LCA) unit process models [6]. Though the focus of these two standards lies within environmental analysis, the reference models should be robust enough to include traditional performance attributes associated with manufacturing systems, e.g. cost, quality, and throughput.

One of the motivating factors for implementing a standard representation for manufacturing information is model storage and curation. In this light, we proposed an open web-based repository to promote data consistency and bridge the research gap between institutions and private sectors [7]. This repository adopts the ASTM E3012-16 standard and uses the Unit Manufacturing Process (UMP) as a formal model for capturing manufacturing data [5]. The UMP captures input and output information as well as various process parameters and can be modeled using data formatting languages such as eXtensible Markup Language (XML) or JavaScript Object Notation (JSON).

This paper focuses on the development of a metric to assess the similarity between such UMP models. We envision that this work will improve the navigability and usability of the proposed repository. Next, we review existing database incarnations for the manufacturing processes, specifically in the context of supplier discovery and process capability-based query.

2.2 Storing process capability information

Currently, there are several databases that store manufacturing information. The CES Selector from Granta Design[1], a commercial database designed for material selection, houses general information on manufacturing processes. Such information includes broad interval ranges of performance indicators, e.g. cost and CO_2 emissions per amount of material processed. Since the manufacturing capabilities of an organization are heavily dependent on their acquired manufacturing assets and resources, commercial databases, such as CES Selector, do not provide much specificity for manufacturing processes. From another perspective, there are a number of supplier discovery and service matching tools, e.g. Alibaba[2], in which it is possible to match process capability with available resources. However, in these tools' current form, such an effort would require significant costs in time to translate this information into a usable form. Also, it is possible that the posted capabilities of individual job shops and manufacturers could be over-claimed and not accurate [8].

In response to these challenges, there have been efforts in providing open and free access to manufacturing process capability information, such as CustomPart.Net[3]. Such databases provide a variety of estimation and manufacturing tools based on tabular information of material and manufacturing process, and supplier information. Examples of available tools include a milling speed and feed calculator, cost estimation for injection molding, and bend allowance calculator for sheet metal. These open sources claim widespread use in industry with thousands of reported estimations per month. In this paper, we use this type of information about manufacturing processes since the data is open, available, and seemingly trustworthy judging by its wide use as well as its use within similar research efforts [9].

2.3 Defining similarity between information models

Measuring the similarity between information models is nothing new. By definition, the similarity between two objects is a function of the commonality and the differences they share [10]. In this paper, we borrow concepts from similarity

[1] https://www.grantadesign.com/products/ces/
[2] https://www.alibaba.com/
[3] http://www.custompartnet.com/

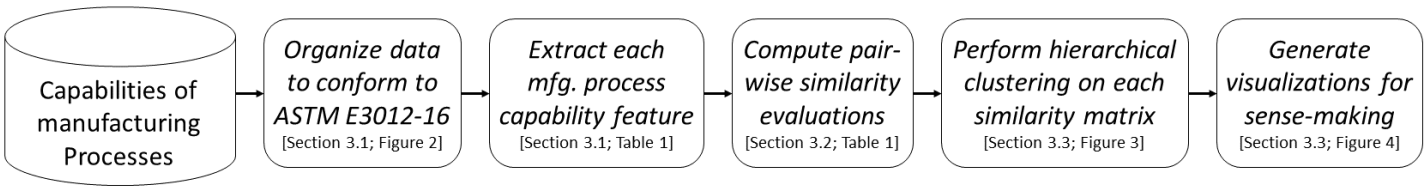

FIGURE 1. Flowchart specifying steps for the analysis performed. Here, we transform tabular data about manufacturing processes into a standard representation, abiding by ASTM 3012-16. We then compare each pair of manufacturing processes based on the computed similarity metrics, perform clustering on the resultant matrix, and generate visualizations as a reporting mechanism to aid in data sense-making.

measures from various applications areas, including biology, information science, and engineering.

In biomedical research, similarity measures have aided in the prediction of functions and interactions between different gene products such as proteins. These measures often utilize semantic similarity techniques due to an abundance of biomedical annotations such as the Gene Ontology [11]. In fact, now a fully adopted method, visual representations of microarray data, which represents this expression-based similarity, facilitates a deeper insight and understanding of the data to practitioners [12].

Likewise, incorporating ontologies is a key factor in building a successful model for representing manufacturing data. One attempt at creating a digital manufacturing ontology is the Manufacturing Service Description Language (MSDL) [13]. MSDL captures the abstract concepts and relationships between manufacturing services at a process, machine, shop, and supplier level. Other examples for assessing the similarity between manufacturing processes can be classified as edge-based counting methods. In these approaches, a taxonomic representation of the process universe, such as the Allen and Todd taxonomy [14], is used to determine distance between nodes in a network, e.g. in [15].

Similar approaches have also been used in the conceptual design phase, wherein only the function-component relationships are considered [16, 17], as well as in detailed design, wherein researchers have used similarity measures to uncover opportunities from existing designs [18, 19]. Others have focused specifically on cellular manufacturing, developing a metric for assessing the similarity of production lines [20]. Another approach used a graph-based metric to assess the similarity of manufacturing-based value chains [21].

Due to a lack of abundant annotations in the manufacturing industry, it can be argued that common ontology measures such as node-based approaches using information content (IC) and edge-based counting methods cannot be applied to current manufacturing data [22]. Edge-based methods using manufacturing process trees are also unreliable due to the uneven distribution of nodes and the inability to quantify the length between each parent and child node. In response, in this paper, we propose a hybrid approach wherein we combine both semantic and numerical information to develop a similarity metric to assess a manufacturing process's capabilities.

3 METHODOLOGY

In this paper, we present work towards a more comprehensive similarity metric for manufacturing processes based on their capabilities, such as achievable tolerance, surface roughness, and batch size. One of the primary challenges lies within the fact that manufacturing-related information housed in databases come in different forms, such as categorical and numerical expressions. As shown in Fig. 1, this section details the steps towards the development of this metric, including (1) data selection and processing, (2) the implementation of several different similarity calculations based on the data quality and nature of each selected capability, and (3) the visualization of the computed metrics to aid in decision making.

3.1 Pre-processing and organizing the data

Choosing the source of data was the first step as accessibility, representation, and uniformity of data were important factors to consider for an intuitive similarity measure. CustomPart.Net was chosen as it is an open-source database that can easily be accessed online and has an abundance of pre-formatted data. The database can also easily be expanded by collaborative use from institutions and private sectors. Access to larger, more specific amounts of data can enhance the existing similarity measures while also allowing for new measures to be introduced.

The standard format of a UMP model was used to capture the chosen dataset. In order to represent a UMP model, XML was chosen as the best format due to its simplicity to use and intuitive structure for human and machine reading.

A script was written using MATLAB that constructed individual XML files for each process from a conglomerate Excel spreadsheet. The raw data was captured in multiple formats and had to be properly sorted. The select features that were specified numerically occasionally came in ranges. These values were captured into upper and lower bounds as seen in the surface finish values in Fig. 2. Manufacturing data is also traditionally captured

```xml
<?xml version="1.0" encoding="utf-8"?>
<UMP name="Milling">
    <ProductProcessInformation name="Shapes_Typical"
        value="">
      <string-array name="Typical" value="">
        <item>Solid: Cubic</item>
        <item>Solid: Complex</item>
      </string-array>
    </ProductProcessInformation>
    <ProductProcessInformation name="Shapes_Feasible"
        value="">
      <string-array name="Feasible" value="">
        <item>Flat</item>
        <item>Thin-walled: Cylindrical</item>
        <item>Thin-walled: Cubic</item>
        <item>Thin-walled: Complex</item>
        <item>Solid: Cylindrical</item>
      </string-array>
    </ProductProcessInformation>
    <ProductProcessInformation name="Materials_Typical"
        value="">
    <ProductProcessInformation name="Materials_Feasible"
        value="">
    <ProductProcessInformation name="Lead_Time_Typical"
        unit="" value="Days"/>
    <ProductProcessInformation name="Lead_Time_Feasible"
        unit="" value="Hours"/>
    <ProductProcessInformation name="Tolerance_Typical"
        unit="in" value="0.001"/>
    <ProductProcessInformation name="Tolerance_Feasible"
        unit="in" value="0.0005"/>
    <ProductProcessInformation name="Max_Wall_Thickness
        _Typical_UB" unit="in" value="40"/>
    <ProductProcessInformation name="Max_Wall_Thickness
        _Typical_LB" unit="in" value="0.04"/>
    <ProductProcessInformation name="Max_Wall_Thickness
        _Feasible_UB" unit="in" value="72"/>
    <ProductProcessInformation name="Max_Wall_Thickness
        _Feasible_LB" unit="in" value="0.04"/>
    <ProductProcessInformation name="Surface_Finish_
        Typical_UB" unit="microinch" value="125"/>
    <ProductProcessInformation name="Surface_Finish_
        Typical_LB" unit="microinch" value="32"/>
    <ProductProcessInformation name="Surface_Finish_
        Feasible_UB" unit="microinch" value="500"/>
    <ProductProcessInformation name="Surface_Finish_
        Feasible_LB" unit="microinch" value="8"/>
    <ProductProcessInformation name="Batch_Size_
        Typical_UB" unit="" value="1000"/>
    <ProductProcessInformation name="Batch_Size_
        Typical_LB" unit="" value="1"/>
    <ProductProcessInformation name="Batch_Size_
        Feasible_UB" unit="" value="1000000"/>
    <ProductProcessInformation name="Batch_Size_
        Feasible_LB" unit="" value="1"/>
    <ProductProcessInformation name="Applications"
        value="">
      <string-array name="">
        <item>Machine Components</item>
        <item>Engine Components</item>
      </string-array>
    </ProductProcessInformation>
</UMP>
```

FIGURE 2. Example of a UMP generated wherein data from CustomPart.Net was organized via the ASTM E3012-16 standard.

in terms of typical and feasible data, where typical data reflects how a process is traditionally used, while feasible represents the physical limitations of the process. These limitations may be achieved at a sacrifice of efficiency in cost, energy, or production speed. Categorical data such as the materials were captured into string arrays for ease of processing later on.

In order to analyze the manufacturing data from any given database, the data can be entered into an Excel spreadsheet following a specific format, and similar UMP files will be generated. MATLAB was again chosen to run the remaining data processing and analysis. Using object-oriented programming methods, a simple "Process" class was created that reflected the format of the UMP by capturing each feature information into individual instance variables of the class.

3.2 Applying similarity metrics to process attributes

The information housed in the manufacturing database can be grouped into two classifications: numerical and categorical. Numerical data is expressed as nominal values or a range of values. Categorical data can be described as a list of attributes falling into a single classification. The treatments of each type of information are further explained below.

3.2.1 Numerical data For numerical information, wherein the values do not deviate by orders of magnitude, we used the Euclidean distance measure as seen in Eq. 1 and normalized it into a simple similarity function, as shown in Eq. 2.

$$D(x,y) = \sqrt{\sum_{i=0}^{n}(x_i - y_i)^2} \tag{1}$$

$$S(x,y) = \frac{\sqrt{x^2+y^2} - \sqrt{(x-y)^2}}{\sqrt{x^2+y^2}} = 1 - \frac{D(x,y)}{norm(x,y)} \tag{2}$$

where $D(x,y)$ denotes the Euclidean distance and $S(x,y)$ denotes the similarity of one entity, x, with respect to another entity, y. The element i refers to the number of elements or dimensions within a category, which was typically just one for this paper.

By applying the norm to Eq. 1, it can be rewritten to yield a result from 0 to 1, where a 1 signifies a perfect match. It is important that every metric is normalized onto the same scale in order to compare different features together in an aggregate equation. This concept will be explored later in the paper.

$$S(x,y) = exp[-\frac{\log(1+|x+y|)}{k}] \tag{3}$$

$$k = \sum_{i=0}^{n}\frac{\log(|x_i - y_i|)}{n} \tag{4}$$

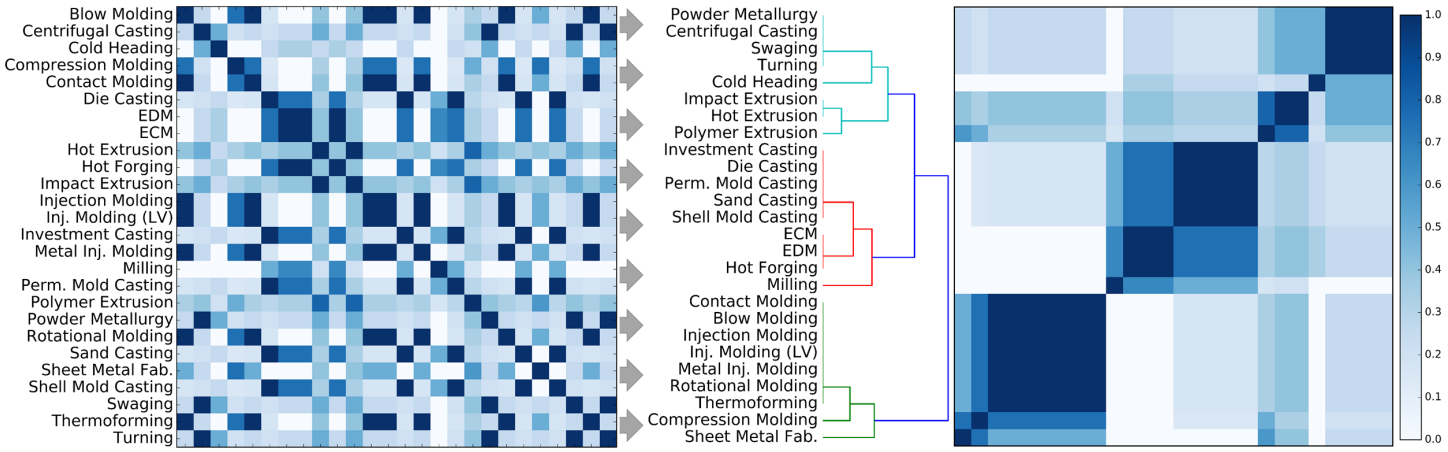

FIGURE 3. Left: example of a square similarity matrix generated. In this case, we are showing the results from a Jacquard index evaluation comparing categorical sets representing typical shapes that each process can produce. Here, we apply a colormap to indicate the level of similarity between two manufacturing processes, wherein the darker the blue denoting a higher similarity. Right: results of hierarchical clustering performed on this similarity matrix. The clusters are illustrated via a dendrogram, where its leaf labels correspond to the rows of the matrix on the right.

TABLE 1. Capability features extracted from CustomPart.Net

#	Feature	Type	Sim. Calc.
1	Shapes	Categorical	Eq. 5
2	Materials	Categorical	Eq. 5
3	Surface Finish	Numerical	Eq. 2
4	Tolerance	Numerical	Eq. 2
5	Max Wall Thickness	Numerical	Eq. 2
6	Applications	Categorical	Eq. 5
7	Batch Size	Numerical	Eq. 3
-	Aggregate Metric	Mixed	Eq. 7

Equation 3 was developed as a result of the batch size feature in the manufacturing data which differed on orders of magnitude. The data ranged from values of 10 to 1,000,000 which made ordinary distance measures such as Euclidean distance ineffective. One method that was attempted was to use the z-score technique to rescale each value based off of standard deviations. However, this method proved ineffective due to the the lack of spread in the data, as most values were captured in magnitudes of 10 rather than in specific quantities. An exponential function was picked as the metric of choice as it can scale values regardless of their magnitude. The exponential of the negative logarithm is a technique used to normalize the distance to a scale of 0 to 1. In order to achieve a similarity score of 0 for perfect similarity between two values, the value inside the logarithm needed to be scaled by 1. Finally, a weight, k, is used to alleviate the spread of the values, such that a large k value would yield a higher similarity score when comparing two values such as 10 and 1,000,000. After testing different k values, the mean of the log of every difference combination was chosen as seen in Eq. 4.

It should be noted that some attributes, such as batch size, were expressed as ranges. These intervals were separated into upper and lower bounds, denoted as UB_n and LB_n, respectively. Then, a similarity measure is applied to every combination of values such as UB_1 to UB_2, UB_1 to LB_2, etc. For simplicity, the largest similarity measure was taken out of the combinations.

3.2.2 Categorical data For categorical information, the Jaccard Index was adopted as a common measure for comparing two sets. As seen in Eq. 5, the similarity between two sets A and B can be taken as the intersection over the union of A and B. This is a simple but effective way to measure similarity for the materials and shapes where the data is evenly distributed among every process.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{5}$$

Table 1 summarizes the manufacturing capabilities that were extracted from CustomPart.Net and the respective similarity metric applied to each. For each features, similarity matrices were computed to represent a total of 325 pairwise comparisons amongst 26 different processes. Next, we explore visualization options for presenting results from applying the various similarity measures.

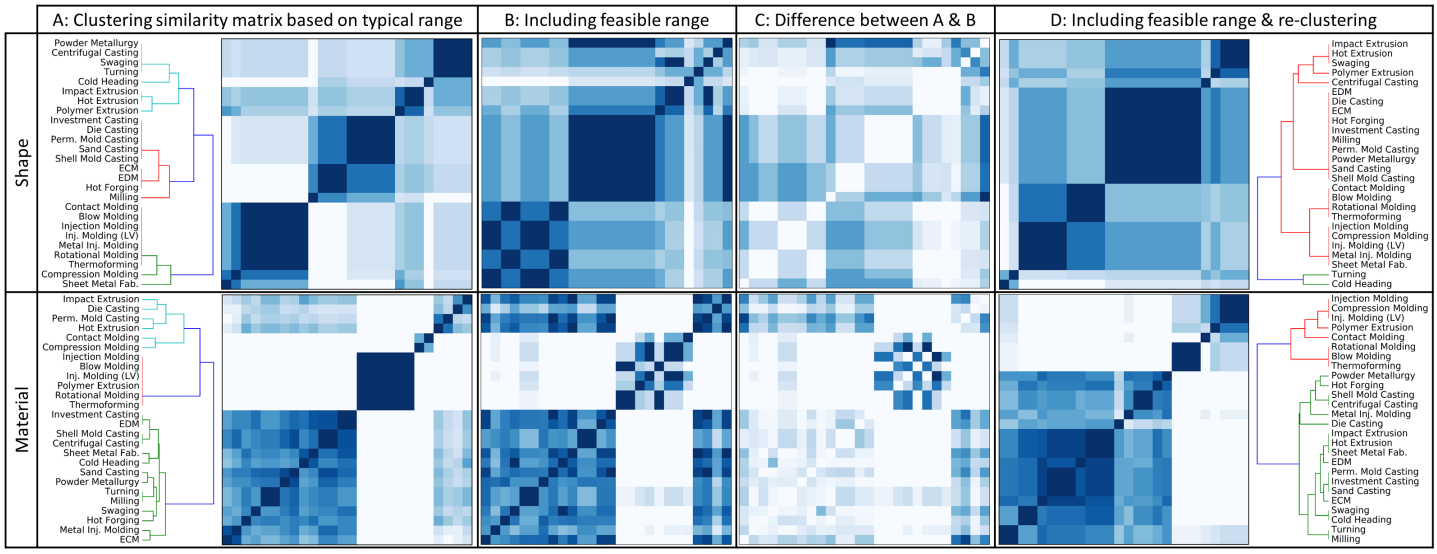

FIGURE 4. Analysis of the effect of expanding the set of typical features with feasible ranges. Here, we consider two of the studied capability features, shape (top) and material (bottom). Column A shows the results from the hierarchical clustering following the same procedure shown in Fig. 3. Column B illustrates the difference in similarity evaluation when including the feasible range. Note that we do not perform any re-clustering at this stage. Column C presents the difference between Columns B & A. Column D shows the shows the results from re-clustering based on Column B.

3.3 Visualizing the similarity matrices

Applying prudent visualization to scenarios with rich analytics allows human decision makers to quickly gain insight into data. This process of gaining insight through internal cognitive processes is commonly referred to as sense-making [3]. Here, we exploit accepted matrix-based visualization to gain deeper insight into the proposed similarity metrics. The primary visualization features are the clustered similarity matrix codified with a color scale, and its accompanying dendrogram.

First, we begin with each computed similarity matrix. An example of one matrix can be seen in Fig. 3 on the left. This matrix represents the values from the similarity evaluation depending on the metric used, wherein the darker the blue denotes the higher value of similarity. This is a square matrix, where the rows and columns represent the same set of processes.

Based on the values in the similarity matrix, we perform hierarchical clustering to identify communities of manufacturing processes based on sharing similar capabilities. To evaluate the distances between resultant clusters, we employ the Voor Hees Algorithm [23], seen in Eq. 6.

$$d(u,v) = max\{dist(u[i],v[j])\} \tag{6}$$

Based on the computed values from Eq. 6, we then build a dendrogram, to visualize the distances between the clusters, which can been seen in Fig. 3 on the right. A longer line translates to a larger distance between the clusters. If there is a vertical line adjacent to the leaf labels, this signifies that the processes are precisely the same based on that specific similarity evaluation. For example, in Fig. 3, *Powder Metallurgy*, *Centrifugal Casting*, *Swaging*, and *Turning* are exactly identical with respect to their capability of producing a typical set of shapes.

4 RESULTS

This section presents some insights drawn out by analyzing the similarity matrices representing various capability features. One of those insights is presented in Fig. 4. Here, we demonstrate some of the challenges associated with dealing with the data. The database provides two different ranges for each feature studied, one that describes the *typical* range of process capability and another broader set that characterizes the *feasible* range.

Figure 4 shows that including the feasible set of attributes affects the specific feature differently. In this case, we consider two capability features, shape and material, both of which are defined sets of strings. Column A presents both similarity matrices clustered solely based on the typical set of attributes (i.e. possible shapes and materials). We then append the sets with the additional attributes from each feasible set, as seen in Column B. Here, we re-calculate the similarity matrix based on the new sets of capabilities but do not re-order the rows. By investigating the difference of Column A and Column B, presented through a simple subtraction process in Column C, we show that, in these cases, the feasible range affects the process communities differ-

TABLE 2. Corpus of attributes for shape and material

Shapes–6 total	Materials–17 total
Flat; Thin-walled: Cubic; Thin-walled: Cylindrical; Thin-walled: Complex; Solid: Cylindrical; Solid: Cubic; Solid: Complex	Alloy Steel; Carbon Steel; Cast Iron; Stainless Steel; Aluminum; Copper; Lead; Magnesium; Nickel; Tin; Titanium; Zinc; Ceramics; Composites; Thermoplastics; Thermosets; Elastomers

ently. This is further enforced by re-clustering the similarity matrix based on the total set of attributes, including both typical and feasible ranges, seen in Column D.

To explain this difference in sensitivity, let us take Electrical Discharge Machining (*EDM*) and Electrochemical Machining (*ECM*) as an example. With respect to shape, as seen in Fig. 4A, *ECM* and *EDM* share identical sets of typical shapes with *Hot Forging*. After appending the original set of possible shapes with the feasible range, this cluster is expanded as seen in Fig. 4D. It now includes processes such as *Die Casting*, *Investment Casting*, and *Milling*. In summary, considering the complete range of feasible shapes, clusters become more inclusive.

If we consider the same processes, *EDM* and *ECM* for the material feature, we find key differences when compared to the shape feature. As seen in Fig. 4A, *EDM* is identical to *Investment Casting* and *ECM* is closest to *Metal Inj. Molding*. However, after appending with the feasible set and re-clustering (Fig. 4D), we see that clusters are less inclusive and often exhibit different characteristics. For example, *EDM* now represents a unique set of capabilities and is closest to *Sheet Metal Fab.*, *Hot Extrusion*, and *Impact Extrusion*.

This observation can be explained by investigating the nature of the data. Table 2 lists each corpus of terms for both the shape and material categories. The materials feature almost 3 times the amount of terms when compared with shapes, i.e. 17 compared to 6. This partly explains why the shape feature has much more inclusive clusters after including the feasible ranges. With fewer possible attributes, the shape feature can be expressed by a reasonable number of possible combinations. Once seeded with more possibilities, e.g. 17 concepts as in the materials feature, the clustering behaves differently.

Figure 5 shows other examples of clustered similarity matrices for selected features, only based on typical ranges of capabilities. Similar to above, the nature of the data drastically affects the "performance" of the hierarchical clustering. For Surface Finish, which is numerical data wherein similarity was computed based on Eq. 2, we see a large number of small clusters that exhibit considerable closeness. This is due to the variability in the data's numerical range but the metric still seems to successfully delineate communities within that range. For Batch Size, we see large clusters due to the the homogeneity of the database infor-

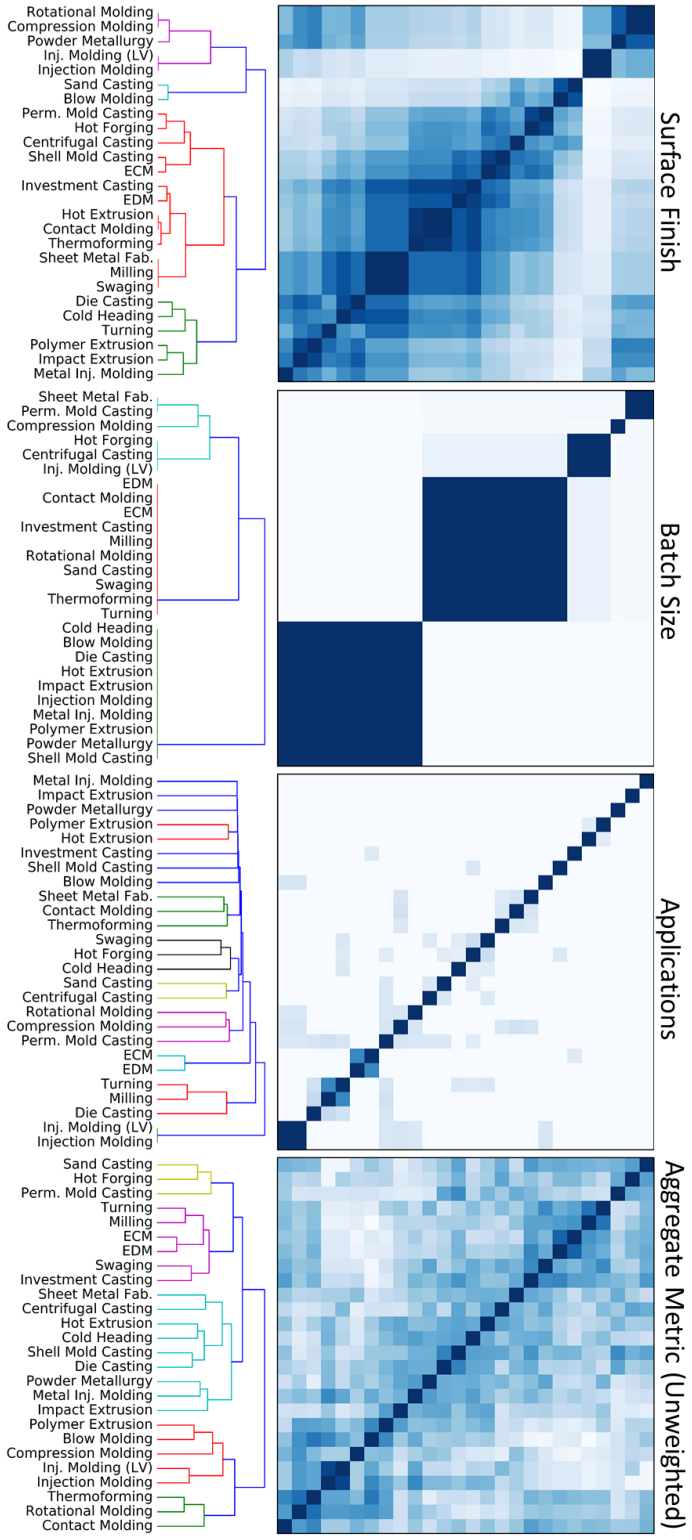

FIGURE 5. Clustering typical surface finish, batch size, applications, and an aggregate metric considering all features.

mation. In contrast, for Applications, which represent a corpus of 72 terms specifying engineering applications (e.g. gears, pipes, and aerospace components), we see few identifiable similar sets of processes due to the heterogeneity of terms.

The significant takeaway of our analysis is that the corpus of concepts for each capability feature has a seemingly large impact on the clustering. This becomes a challenge when developing an aggregate similarity metric that takes into account all capability features studied. The last matrix shown in Fig. 5, reflects the similarity calculations of every feature combined using an aggregate measure as seen in Eq. 7, where k is the weight of the feature i.

$$S(x,y) = \sqrt{\sum_{i=0}^{n} k_i (S_i)^2} \tag{7}$$

Without weighting the features within the overall distance measurement, we can see that the clustering algorithm does not identify distinct and tight clusters. Instead, we see a number of "loose" clusters, wherein it is difficult to discern similar processes based on the full set of capability features.

Based on these observations, we describe limitations of our work to motivate future work in the next section. This description will help account for these discrepancies within the data to specifically address the development of an aggregate similarity metric for manufacturing processes.

5 LIMITATIONS OF WORK

Limitations of our approach are tightly related to the nature of the data applied in this paper. Primary issues include (1) data granularity, (2) the heterogeneity of scales for the numerical data, (3) lack of understanding of the inter-relationships for various combinations of capability features, and (4) the scope and context of the data.

As shown in the previous section, the metrics that we applied to the data perform differently, sometimes poorly, based on the granularity of the data. When using semantic metrics such as the Jaccard Index, the distribution of terms and number of unique terms may lead to insubstantial measurements. The Application feature was a clear example of such a scenario. The dataset used 17 unique material types and 72 different application types. Although more data types may appear to benefit the similarity measure, with only 26 processes, the distribution of application types to processes was sparse, with many processes having unique applications. This also leads to issues in syntactic matching as similar words such as aviation and aerospace will not be matched even if they are semantically related. Currently, our method here does not address this issue. We do not treat the semantic similarity measures any differently based on the number of terms in a feature's corpus.

Another challenge is the significant difference in scale for numerical data, which could affect the aggregate similarity metric. For example, differences between batch size capability in the dataset differ on orders of magnitude, while differences in achievable tolerances are not as far apart, yet have critical importance (in real manufacturing scenarios) even with minor differences. In our approach, we use traditional feature scaling to curb these effects. However, this issue seems to pop up in the aggregation of multiple metrics with one another. A smarter approach to overcoming scale heterogeneity in the data is needed.

In addition, our method does not consider the interdependencies between various combinations of capability features. For example, material and shape characteristics of a product have significant interplay judging the feasibility of a real build. Down-selecting from a set of materials influences the feasible range of shapes, and vice versa, even before selecting a set of available processes. From initial experimentation, the interplay of correlated features also have an effect on the generated clusters of material for a weighted aggregated metric. The implications of assigning these weights have a seemingly significant impact on the clustering result. This limitation presents future research questions as to how to emphasize particular aspects of manufacturing capabilities to better inform process selection.

Lastly, the developed metrics are constrained based on the dataset that was used. The data does not model any specific design case, but rather general design parameters that are purely process specific. Empirical methods of data collection have been proven to accurately model and predict important aspects of manufacturing processes such as unit energy consumption [24]. The addition of such data can significantly enhance our current methods while also increasing their utility as energy is a major factor in calculating cost and environmental effect. Features from the dataset such as the shape also lack the depth to accurately classify specific products with complex designs. There have been assessments of shape signatures through spacial functions and histograms that could be applied to our metrics [25]. Analyzing the categorical and numerical comparisons of certain features will require more data that will require expanding beyond our current dataset.

6 CONCLUSION AND FUTURE DIRECTIONS

This paper highlights efforts towards the development of a similarity metric for manufacturing processes. The primary goal is to aid in decision-making in the context of supplier discovery, e.g. given a set of design requirements, and define the available set of processes and viable alternatives based on similar capability-based characteristics. Here, we review lessons learned from our experimentation and present future directions to address the limitations of our work.

A primary research direction is the inclusion of a weighting scheme for individual similarity evaluations of capabilities based on feature inter-dependencies or human preference. Inter-dependencies of capability features could also be heavily

domain-specific. For example, requirements for material selection and tolerance specification vary greatly between consumer products and aerospace applications.

Furthermore, in this work, we have not validated any of the identified clusters of manufacturing processes with human judgement and expert experience. Such tacit knowledge can be efficiently captured in description logic and then formally expressed in ontologies. We have yet to make the connection between our fully automated approach and some flexibility for human-operated tuning. One way of validating the clusters is to elicit expert advice for designing process plans on several simple assemblies. These assemblies would have characteristics similar to the features studied in the paper, including shape, material, and required tolerances. It remains unknown, however, if it would be possible to map these expert decisions to the clusters of processes in order to decide on a particular set of weights for the aggregate metric. Considering various efforts in storing such decisions within knowledge bases, there are some research opportunities for developing a more semi-automated technique for assessing similarity.

Finally, a data-driven metric to holistically evaluate the similarity of two manufacturing processes must be generalizable enough to work across different databases including differences in the granularity of information. Here, we explore the idea using very general tabular data from an on-line resource. However, if the manufacturing information was classified within different features or seeded with significantly more information, our similarity evaluation, in its current form, would not function properly. It is critical to develop a similarity measurement technique that is agnostic of a particular database and promotes flexibility in its tuning and eventual use.

DISCLAIMER

No approval or endorsement of any commercial product by NIST is intended or implied. Certain commercial equipment, instruments or materials are identified in this report to facilitate better understanding. Such identification does not imply recommendations or endorsement by NIST nor does it imply the materials or equipment identified are necessarily the best available for the purpose.

ACKNOWLEDGMENT

This work was partly funded by the Summer Undergraduate Research Fellowship (SURF) Program at NIST.

REFERENCES

[1] Manyika, Chui, B. B. D. R. B., 2011. Big data: The next frontier for innovation, competition, and productivity. Tech. rep., McKinsey Global Institute.

[2] Lee, J., Lapira, E., Bagheri, B., and Kao, H.-a., 2013. "Recent advances and trends in predictive manufacturing systems in big data environment". *Manufacturing Letters,* ***1***(1), pp. 38–41.

[3] Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G., 2008. "Visual analytics: Definition, process, and challenges". In *Information visualization*. Springer, pp. 154–175.

[4] ISO 20140-1, 2012. Automation systems and integration - environmental and energy efficiency evaluation method for manufacturing systems - part 1: Overview and general principles. Standard, International Organization for Standardization, Geneva, Switzerland.

[5] ASTM E3012-16, 2016. Standard guide for characterizing environmental aspects of manufacturing processes. Standard, ASTM International, West Conshohocken, PA, USA.

[6] Meinshausen, I., Müller-Beilschmidt, P., and Viere, T., 2014. "The ecospold 2 formatwhy a new format?". *The International Journal of Life Cycle Assessment*, pp. 1–5.

[7] Bernstein, W. Z., Mani, M., Lyons, K. W., Morris, K. C., and Johansson, B., 2016. "An open web-based repository for capturing manufacturing process information". In Proceedings of the ASME 2016 IDETC/CIE, American Society of Mechanical Engineers.

[8] Koh, T. K., Fichman, M., and Kraut, R. E., 2012. "Trust across borders: buyer-supplier trust in global business-to-business e-commerce". *Journal of the Association for Information Systems,* ***13***(11), pp. 886–922.

[9] Weissman, A., Petrov, M., and Gupta, S. K., 2011. "A computational framework for authoring and searching product design specifications". *Advanced Engineering Informatics,* ***25***(3), pp. 516–534.

[10] Lin, D., 1998. "An information-theoretic definition of similarity". In Proceedings of the 15th ICML, Vol. 98, International Conference on Machine Learning, pp. 296–304.

[11] Pesquita, C., Faria, D., Falcao, A. O., Lord, P., and Couto, F. M., 2009. "Semantic similarity in biomedical ontologies". *PLoS Computational Biology,* ***5***(7), p. e1000443.

[12] Saldanha, A. J., 2004. "Java treeviewextensible visualization of microarray data". *Bioinformatics,* ***20***(17), pp. 3246–3248.

[13] Ameri, F., and Patil, L., 2012. "Digital manufacturing market: a semantic web-based framework for agile supply chain deployment". *Journal of Intelligent Manufacturing,* ***23***(5), pp. 1817–1832.

[14] Todd, R. H., Allen, D. K., and Alting, L., 1994. *Manufacturing Processes Reference Guide*. Industrial Press Inc.

[15] Ramanujan, D., Bernstein, W. Z., Benjamin, W., Ramani, K., Elmqvist, N., Kulkarni, D., and Tew, J., 2015. "A framework for visualization-driven eco-conscious design exploration". *Journal of Computing and Information Science in Engineering,* ***15***(4), p. 041010.

[16] McAdams, D. A., and Wood, K. L., 2002. "A quantitative similarity metric for design-by-analogy". *Journal of Mechanical Design,* ***124***(2), pp. 173–182.

[17] Fernandes, R. P., Grosse, I. R., Krishnamurty, S., Witherell, P., and Wileden, J. C., 2011. "Semantic methods supporting engineering design innovation". *Advanced Engineering Informatics,* ***25***(2), pp. 185–192.

[18] Avramenko, Y., and Kraslawski, A., 2006. "Similarity concept for case-based design in process engineering". *Computers & Chemical Engineering,* ***30***(3), pp. 548–557.

[19] Witherell, P., Grosse, I. R., Krishnamurty, S., and Wileden, J. C., 2013. "Aiero: An algorithm for identifying engineering relationships in ontologies". *Advanced Engineering Informatics,* ***27***(4), pp. 555–565.

[20] Shafer, S., and Rogers, D., 1993. "Similarity and distance measures for cellular manufacturing. part ii. an extension and comparison". *The International Journal of Production Research,* ***31***(6), pp. 1315–1326.

[21] Zhu, Z., Morrison, G., Puliga, M., Chessa, A., and Riccaboni, M., 2015. "The similarity of global value chains: A network-based measure". *arXiv:1508.04392*.

[22] Resnik, P., et al., 1999. "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language". *Journal of Artificial Intelligence Research,* ***11***, pp. 95–130.

[23] Müllner, D., 2011. "Modern hierarchical, agglomerative clustering algorithms". *arXiv:1109.2378*.

[24] Kara, S., and Li, W., 2011. "Unit process energy consumption models for material removal processes". *CIRP Annals-Manufacturing Technology,* ***60***(1), pp. 37–40.

[25] Cardone, A., Gupta, S. K., and Karnik, M., 2003. "A survey of shape similarity assessment algorithms for product design and manufacturing applications". *Journal of Computing and Information Science in Engineering,* ***3***(2), pp. 109–118.