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

IDETC/CIE 2018 - 86300

AUTOMATIC CLUSTERING OF SEQUENTIAL DESIGN BEHAVIORS

Molla Hafizur Rahman
Department of Mechanical Engineering
University of Arkansas, Fayetteville, AR

Michael Gashler
Department of Computer Science and Computer Engineering
University of Arkansas, Fayetteville, AR

Charles Xie
Concord Consortium
Concord, MA

Zhenghui Sha*
Department of Mechanical Engineering
University of Arkansas, Fayetteville, AR

ABSTRACT

Design is essentially a decision-making process, and systems design decisions are sequentially made. In-depth understanding on human sequential decision-making patterns in design helps discover useful design heuristics to improve existing algorithms of computational design. In this paper, we develop a framework for clustering designers with similar sequential design patterns. We adopt the Function-Behavior-Structure based design process model to characterize designers' action sequence logged by computer-aided design (CAD) software as a sequence of design process stages. Such a sequence reflects designers' thinking and sequential decision making during the design process. Then, the Markov chain is used to quantify the transitions between design stages from which various clustering methods can be applied. Three different clustering methods are tested, including the K-means clustering, the hierarchical clustering and the network-based clustering. A verification approach based on variation of information is developed to evaluate the effectiveness of each method and to identify the clusters of designers who show strong behavioral similarities. The framework is applied in a solar energy systems design problem – energy-plus home design. The case study shows that the proposed framework can successfully cluster designers and identify their sequential decision-making similarities and dissimilarities. Our framework can support the studies on the correlation between potential factors (e.g., designers' demographics) and certain design behavioral patterns, as well as the correlation between behavioral patterns and design quality to identify beneficial design heuristics.

Keywords: Cluster analysis, Design thinking, Markov chain, Design process, Sequential Decision Making, Function-Behavior-Structure.

1 INTRODUCTION

Engineering systems design is a series of interrelated operations that is driven by designers' decisions. Systems design decisions are sequential rather than concurrent optimization strategies [1]. While designers are involved in a design process, they iteratively and sequentially make decisions to explore and exploit the design space in order to improve their designs. As a consequence, sequential decision-making has significant impacts on the quality of design outcomes and the resources needed to achieve the outcomes. A deeper understanding on designers' sequential decision-making behaviors is critical to the discovery of generalized design processes and heuristics that can, in turn, be used to facilitate design process and enhance design automation.

Sequential decision making is an essential component of de-

*Corresponding author: zsha@uark.edu

sign thinking. Dym et al. [2] defines design thinking as a complex process of inquiry and learning that designers perform in a systems context, *making decisions as they proceed* and often working collaboratively. A deeper understanding of designers' sequential decision-making behaviors, especially their patterns, is critical to advancing artificial intelligence in engineering design, for example, by encoding human intelligence in many computational design frameworks.

However, modeling design decision-making is scientifically challenging because human decisions are the result of a mental process that is hidden, implicit, and sometimes tacit [3]. Such a challenge is even more significant in a systems design context that consists of a large number of coupling design variables. To address this challenge, we adopt a data-driven approach and use unsupervised clustering methods to mine designers' sequential design patterns.

The **overall objective** of this study is to establish a framework that automatically identifies and clusters sequential design behavioral patterns of a group of designers. To achieve this objective, the FBS-based design process model is adopted to characterize designers' action sequence logged by computer-aided design (CAD) software as a sequence of design process stages. Then, the Markov chain is used to quantify the transitions between design stages from which various clustering methods can be applied. We adopted three different clustering methods including the K-means clustering, the hierarchical clustering and the network-based clustering. Finally, a verification approach based on information theory is applied to evaluate the effectiveness of each method and to identify the clusters of designers who show strong behavioral similarities. Our study is motivated and driven by answering the following two questions:

- What are the sequential design behavioral patterns that most designers would follow in systems design?
- If designers behave similarly in sequential design making of time domain, would their behaviors quantified in frequency domain are also similar?

This paper is organized as follows. In Section 2, we present relevant literature. In Section 3, the framework of identifying automatic clustering design behavioral patterns is introduced and discussed. The data collection and experiment procedure are described in Section 4. In Section 5, the results are discussed and we conclude the paper with closing thoughts and future work in Section 6.

2 BACKGROUND AND LITERATURE REVIEW

In this section, we review the relevant studies about sequential decision-making in the engineering systems design field. Also, to understand the sequential decision making in a design process, it is important to have a model at the first place to characterize the design process such that a coding protocol can be used to encode and computationally extract the sequential design decisions. Different models hold different assumptions. Identifying an appropriate design process model is thus a critical step to answering the research questions aforementioned. Therefore, in Section 2.2, we provide a literature review on existing studies of design process models.

2.1 Sequential Decision Making in Engineering Design

In engineering design, different models and theories have been used to study designers' sequential decision making. To facilitate the development of sequential guidelines and shorten design cycle, Smith and Eppinger [4] developed a sequential iteration model. The author adopted the design structure matrix that sequences the design task in an optimized way. Then the expected execution time for engineering development project can be predicted. Yukish et al. [5] developed a formal model that describes the sequential decision process by addressing the issues of low fidelity model and high fidelity model in engineering design. The authors also showed how low fidelity model can be coupled with high fidelity model for ease of detail modeling. In order to understand the sequential decision-making behaviors in design under competition, Sha et al. [6] developed a model that integrates game theory with Bayesian optimization (BO). Using design crowd sourcing as an example, the authors adopted the non-cooperative game and Wiener process-based BO to estimate designers' trade-off preferences in design while two players compete for wining a monetary award.

Markov chain is a widely used technique in design area to understand designers' sequential decisions. Yu et al. [7] applied the first-order Markov chain and Function-Behavior-Structure (FBS) ontology to explore the effect of the design knowledge and experience on design patterns in a parametric design environment (PDE) and geometric modeling environment (GME). Later second-order Markov chain was implemented to the same purpose. From the study, it is found that designers exhibit more design patterns in PDE than GME. Kan and Gero [8] also used the first-order Markov chain to compare designers' behaviors in three different design domains: architectural design, software design and mechanical design. In a recent study, McComb et al. [9] find that the first-order Markov chain better represents designers' sequential decisions than higher-order Markov chain in configuration design problems. Existing studies using Markov chain are mainly focused on identifying designers' behaviors at an aggregate level. However, each designer may have different sequential behavioral strategies. Understanding of individual design strategies is essential to design research in many aspects, such as informing better structure of design teams, developing customized CAD software, fostering personalized learning, and identifying design experts vs. novice, which is an important topic of design knowledge acquisition and management. Therefore, in this

paper we aim to address the gap by analyzing designers' sequential decision-making behaviors at the individual level to study dissimilarities and then cluster similar behaviors for identifying potential design behavioral patterns.

2.2 Design Process Model

Design process model is a central element of design methodologies [10]. Depending on the number of stages that a design process can be divided into, various design process models are developed. Asimow [11], Darke [12], March [13] proposed different design process models that all have three stages. Though their models have the same numbers of design stages, the definition, term, and functionality of the design stages are different from each other. Pahl and Betiz [14] identified four main stages: clarification of the task, conceptual design, embodiment design and detail design. Design Council [15] also introduced a four-stage divergent-convergent model, known as double diamond model, which consists of discover, define, develop, and deliver. Howard [16] developed a framework that contains 23 design models mainly from mechanical engineering. In this framework, he identified the similarities among the design process models and mapped the process models to a six-stage model: establishing a need, analysis of task, conceptual design, embodiment design, detailed design, and implementation.

The design process research is often an ontology study, which aims to establish a vocabulary of knowledge representation [17] of design process via logical theory. A typical ontology for the formalization of knowledge about a design process is the FBS ontology [18]. The FBS is constructed with three classes of ontological variables: Function (F), Behavior (B), and Structure (S). Later, two additional variables are added for better representation of the design process: Requirements (R) and Descriptions (D). Umeda et al. [19] introduced a different model but also called FBS (Function, Behavior, State) to clarify the distinction between function and behavior. Later, Deng et al. [20] added working environment to the model and proposed Function-Behavior-Structure-Environment (FEBS) model. Similarly, many other FBS variants, such as B-FES model, Onto-FaBeS model and RFBS model, are developed in the past few years. Though a lot of frameworks are developed, FBS seems to be a universal framework for different design environment and process [21]. This is also the motivation for us to adopt the FBS model in this study for characterizing the design process.

3 OVERALL METHODOLOGY

3.1 General Approach

In this section, we present our approach to clustering designers' sequential decision-making behaviors. As shown in Figure 1, we use several data types which are converted from one to another. Each data type is described as below:

- *Design action data*: In this study, design actions are defined as the design-related operations used in a CAD environment, for example, adding a new component or changing the size of a component.
- *Design process data*: Design process data is transformed from design action data by a design process model, e.g., the waterfall model or the spiral design model, etc. The design process data has a reduced dimensionality as compared to the design action data depending on the number of processes defined by a design process model.
- *Design behavioral data*: This the data generated from design behavior models by taking design process data as the input. The resulting data characterizes and quantifies design behavioral features. For example, if using Markov chain to study the sequential decision-making behaviors, the design process data will be converted to the transition probability matrix (see Section 5.1 for details), which is regarded as the design behavioral data.

Once the design behavioral data is obtained, different clustering methods can be applied to group designers with similar behavioral patterns. The optimal number of clusters can be determined from a standalone method, e.g., using elbow plot [22] in K-means clustering, or sometimes the clustering methods can automatically determine the best number of clusters. Since different clustering methods usually produce different clustering re-

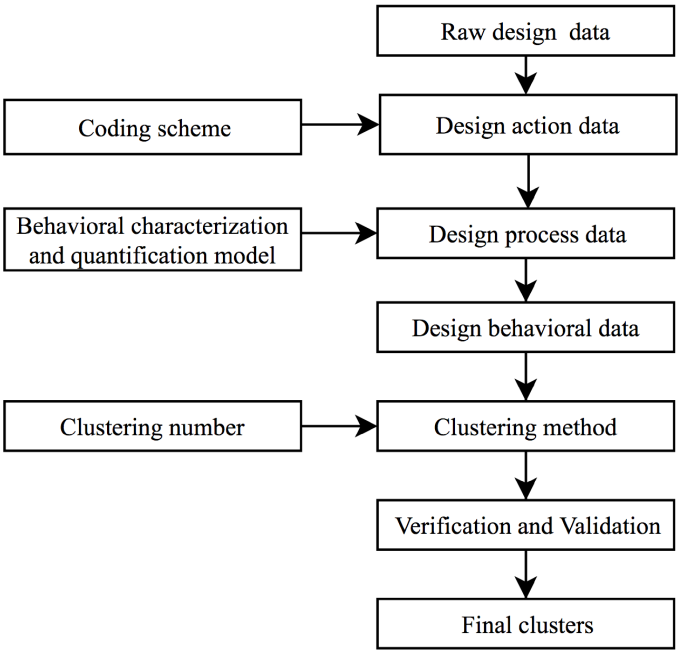

FIGURE 1: The approach to automatically clustering design behaviors

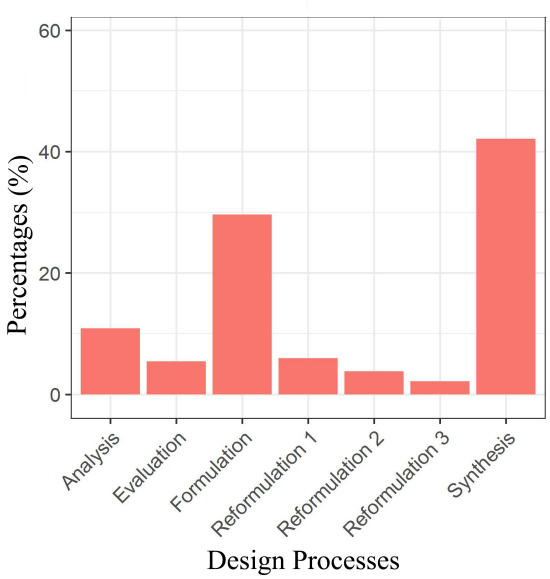

FIGURE 2: Design process stages distribution of designer A10

sults, it is important to verify the results from different methods. Therefore, a verification approach is needed to assure the correctness and the quality of outputs. It's worth noting that each of the components in Figure 1 can be programmed and seamlessly connected to turn the approach into an automatic clustering tool. In the following subsections, we present the details for each step.

3.2 Characterizing Sequential Decisions Using Markov Chain

In this study, the first-order Markov chain [23] is adopted to characterize a design process transitioning from one stage to the another. A Markov chain is a stochastic process in which a system transitions between a finite numbers of discrete states. The traditional definition of Markov chain is regulated by the Markov property – the future state of the process is solely based on its present state. This refers to the first-order Markov chain model [24]. Higher-order Markov models can be developed assuming the next state depends on the current state as well as some number of past states [9]. To define a discrete time Markov chain, we need three components:

- *State space: a finite set S of possible states of the system.*
- *Transition probabilities: a function* $\pi: S \times S \rightarrow R$ *such that* $0 \leq \pi(a,b) \leq 1$ *for all* $a,b \in S$ *and* $\sum_{b \in S} \pi(a,b) = 1$ *for every* $a \in S$.
- *Initial distribution: a function* $\mu : S \rightarrow R$ *such that* $0 \leq \mu(a) \leq 1$ *for every* $a \in S$ *and* $\sum_{a \in S} \mu(a) = 1$

In order to use Markov chain to study the sequential decision-making in design, some treatments are needed to adapt

TABLE 1: The FBS-based design process model

Name of the process	Design process
Formulation	$R \rightarrow F$ & $F \rightarrow Be$
Synthesis	$Be \rightarrow Bs$
Analysis	$S \rightarrow Bs$
Evaluation	$Bs \rightarrow Be$
Reformulation 1	$S \rightarrow S$
Reformulation 2	$S \rightarrow Be$
Reformulation 3	$S \rightarrow F$
Documentation	$S \rightarrow D$

the concepts of Markov chain. While designers explore a design space, design actions performed at different time spots may correspond to the same design process stage. The sequence of how the design space is explored (design action space) can, therefore, be mapped to a design process (design thinking space), where the Markov chain can be established to model the sequential decision making as a time series of design stages. In such a configuration, the system states in the Markov chain corresponds to design process stages, and the 'system' is, therefore, the sequential design thinking being studied. To support the mapping of design actions to design process stages, a coding scheme (see Section 4.3) is developed based on the FBS-based design process model.

Based on the five FBS ontological variables mentioned in Section 2.2, such a design process model consists of eight process stages: *Formulation*, *Analysis*, *Evaluation*, *Synthesis*, *Documentation* and *Reformulation* 1, 2 and 3. Table 1 defines how these design process stages are derived from FBS ontology. *Formulation* transforms Requirement (R) into Function (F) and from Function to Expected Behavior (Be). *Synthesis* generates and tunes Structure based on the Expected Behavior. *Analysis* is defined as the process which is generated from Structure (S). Evaluation is the comparison between the Expected Behavior and the behavior enabled by the actual structure (Bs). The design process that transitions from Structure is called *Reformulation*. Depending on which state the process transitions to, three different process stages can be defined. *Reformulation* 1 is the process transitioning from one structure to a different structure. *Reformulation* 2 describes the transitions from Structure to Expected Behavior; and *Reformulation* 3 is the process from Structure to Function. *Documentation* (D) is the description of the whole process.

Motivated by the second research question, in addition to using Markov chain to study the sequential patterns in design process, we also investigate the design process in frequency do-

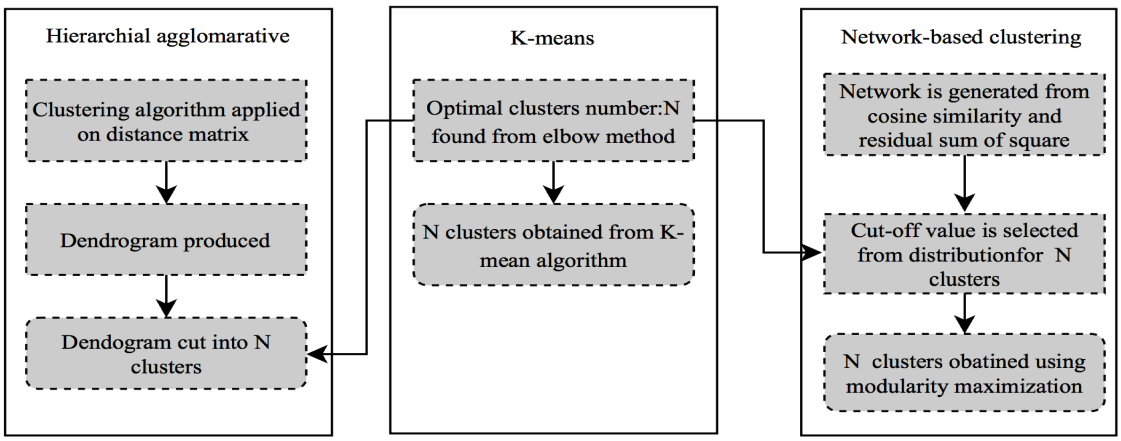

FIGURE 3: Procedure of clustering methods and the selection of optimal clustering numbers for cross comparison

main, i.e., how frequent each of the design stages is utilized by designers in the whole design process. An example of one designer's distribution of design process stages using FBS model is shown in Figure 2. It indicates his/her design utility consists of *Formulation* and *Synthesis*.

Both the transition probability matrix of Markov chain and the distribution of design process stages can be converted to vectors that quantify the features of design behaviors, from which different clustering methods can be applied. For example, the $N \times N$ transition probability matrix generated from first-order Markov chain of one designer can be converted to a $N^2 \times 1$ vector, and a designer's design stages distribution can be converted to a $N \times 1$ vector, where N is the number of stages in a design process model. For n designers, respective $N^2 \times n$ or $N \times n$ matrices will be formed. In this paper, we perform clustering methods on both matrices to analyze designers' sequential decision-making in both time domain and frequency domain.

3.3 Clustering Methods

The goal of a clustering method is to divide data into a meaningful and useful groups based on their similarities [25]. In the field of engineering design, clustering has been used in many applications and various clustering methods e.g., partition-based clustering [26], shape-based clustering [27], hierarchical clustering [28], density-based clustering [29], network-based clustering [30], etc. have been adopted. In this paper, we adopt K-means, hierarchical, and network-based clustering methods to study the differences and similarities of designers' sequential design behaviors. These three different methods are chosen as representatives from three different categories of clustering: hard clustering, flat clustering and network clustering [31], that covers most commonly used clustering methods. A brief description of each method is summarized as follows.

3.3.1 K-means clustering K-means is one of the most popular clustering methods for partitioning dataset into distinct, non-overlapping clusters. The goal is to partition the dataset into K clusters such that the total within-cluster variation, summed over all the clusters, is minimum [32]. There are many ways to define the within-cluster variation. The most commonly used method is squared Euclidean distance. Since K-means requires the number of clusters as input, a separate algorithm is often needed to determine the optimal number of clusters. In this paper, the elbow plot method [22] is used to help make decisions on choosing the number of clusters. For fair comparison across different methods, the number of clusters obtained from the elbow plot method is also used to guide the implementation of the other two clustering methods (see Figure 3), introduced as follows.

3.3.2 Hierarchical clustering Different from K-means clustering method, hierarchical clustering does not require the numbers of cluster as input initially. Instead, it produces a tree-based representation of the observation, called dendrogram. In this study, we adopt the commonly used agglomerative clustering algorithm [32] to generate this dendrogram. This algorithm starts with considering each data point as an individual cluster. The two clusters that are most similar to each other conjugate to one cluster. This process iterates and not stop until all the observations create one group and complete the dendrogram. After the dendrogram is obtained, researchers can cut it into the desired number of clusters.

3.3.3 Network-based clustering In addition to the two clustering methods above, we also develop a network-based clustering approach based on network community detection tech-

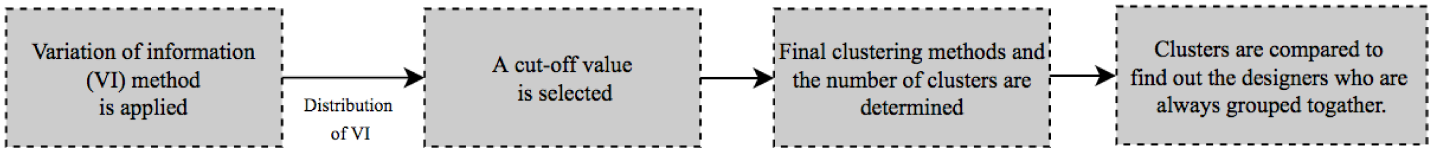

FIGURE 4: Verification of clustering results using variation of information

nique [30]. In this method, a similarity network of designers is first constructed, in which, nodes represent designers and edges represent the similarity between designers. In this study, the residual sum of squares (RSS) [33] and cosine similarity (CS) [33] are used as the similarity metrics. The RSS calculates the sum of the squared differences between the behavioral vectors of two designers, and CS returns the cosine angle between two vectors. Based on their measurement, a similarity matrix can be generated and its elements indicate the similarity between every pairs of designers. In order to retain the strong similarities only, a threshold value is selected to binarize the similarity network. Once the network is ready, different network community detection algorithms can be applied. We utilize the most popular and robust method [34], modularity maximization algorithm [35] to cluster the network. Since the algorithm will automatically cluster the network into an optimized number of clusters, no predetermined number of clusters is needed. To enable the comparison between the three clustering methods, we trial and error the threshold value of similarities (i.e., the RSS and the CS values) until the number of clusters in the network matches the one obtained from K-means elbow plot. See Figure 3 for the whole process and the connection between the network-based and K-means clustering methods.

3.4 Verification & Validation

Since each clustering method produces its own cluster results, for verification purpose, we compare the clustering methods to verify the results. VI is an information-theoretical type of measurement which has been recently found very useful when comparing clustering methods. [36]. VI measures the information lost and gained when it changes from one cluster to another. The lower a VI value is, the better is the partial agreement between two cluster. After obtaining the VI values for each pair of the clustering methods, the methods that have larger partial agreement can be identified, and the designers who have been always grouped together can be found and similar behavioral patterns can be mined from the data. Figure 4 shows the entire procedure. In the following sections, we apply our approach to cluster designers' sequential decision-making behaviors in a solar energy systems design project.

4 CLUSTERING DESIGN BEHAVIORS IN SOLAR ENERGY SYSTEM DESIGN - A CASE STUDY

In this section, we first give a brief description of the design problem. Next, we introduce our experiment procedure for data collection and finally, we present the collected data and FBS-based coding scheme.

4.1 The Design Problem

The design problem in this case study is to build a solarized energy-plus home for a client in Dallas. See an illustrative example in Figure 5. The design objective is to maximize the annual net energy (ANE). The budget for the house is $200,000. In addition to the budget, several design requirements need to be satisfied, as shown in Table 2. The house should have a minimum height of 2.5 m, and the roof must be pitched. The building needs to have at least four windows and one door. The solar panel must be placed on the roof. There are also size constraints on the window, door and the distance between roof ridge and solar panels.

This project is a systems design problem that involves many components (e.g., windows, roof, solar panel, etc.), many design variables (e.g., the number of solar panels, the cell efficiency of

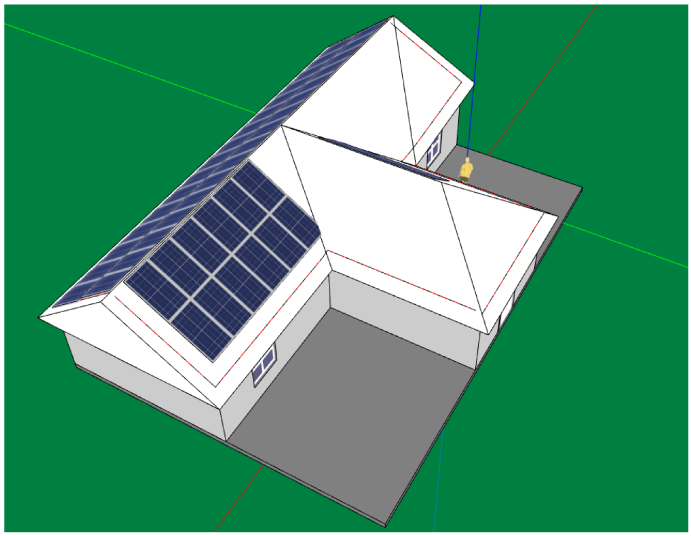

FIGURE 5: An illustrative example of the energy-plus home design project

TABLE 2: The design requirements

Item	Requirements
Story	1
Roof style	Pitched
Number of windows	≥ 4
Size of window	$\geq 1.44\ m^2$
Number of doors	≥ 1
Size of door (Width $\times$ Height)	$\geq 1.2\ m \times 2\ m$
Height of wall	$\geq 2.5\ m$
Solar panel placement	On roof only
Distance between ridge to solar panel	≥ 0

solar panel, etc.), and complex coupling relations among the variables. Therefore, the design space is very large. This is why the requirements and the constraints are developed to reduce designers' action space to a manageable level.

In design problem, designers make trade-off decisions. For example, there is no restriction on the area of the house. But if the area is too small, designers will not be able to place enough solar panel on the roof. As a result, the ANE will be insignificant. On the other hand, if the area is too large, the cost may exceed the budget. So, designers follow their own strategies during the design process to sequentially make decisions guiding the exploration and exploitation of design space so as to improve the ANE as much as possible.

4.2 Experiment Procedure

In order to collect the design action data, a human-subject field experiment [37] is conducted. The energy-plus home design project is performed based on Energy3D – a full-fledged computer-aided design (CAD) tool for solar energy systems [38]. Energy3D has built-in modules of engineering analysis, science simulation and financial evaluation. This ensures the collection of inter-stage design iteration data, e.g., how designers make decisions on a scientific basis (e.g., the ANE analysis results) and economic considerations (e.g., the overall cost), without disrupting the design process and designer's thoughts. Energy3D can automatically log and sort all user actions, at an extremely fine-grained level. All these features enable us to collect high-fidelity data which reflects designers' rational behaviors.

Total 38 people, including both students and faculty members from the University of Arkansas participated in the experiment in five sessions. The participants come from different departments, and the demographics is diverse[1] . The participants are indexed based on which session they were in and which laptop they used. We use letters A to E for the session names, thus A02 means the participant was in Session A and sit in laptop #2.

Each session consists of two phases: pre-session and in-session. The pre-session is thirty minutes long and allotted for participants to practice Energy3D. The design of this pre-session is to account for the learning curve of humans. The data generated in pre-session was not be used for analysis. To further mitigate the learning effects, a tutorial on key operations and terminologies of Energy3D is provided to make sure the participants are familiar with the software environment[2]. At the end of the pre-session the participants will be guided to transition to in-session phase. In-session phase lasts about one hour and half. The design statement and the design requirements are provided at the beginning of this session. A record sheet is provided for participants to record the ANE and cost whenever they iterate their designs.

Monetary rewards are provided at the end of the session to incentivize the participants to explore and exploit the design space as much as possible. The participants are rewarded based on the amount of time they have spent as well as the quality of their final design outcomes, which are related to the ANE and construction cost.

4.3 Data Collection and the FBS-based Coding Scheme

Energy3D logs every performed action and intermediate artifacts (as Energy3D files) every 2 seconds [39]. In our experiment, 220 intermediate files are collected on average and the action log file contains on average 1500 lines of data. The action log file is saved in JSON format and includes time-stamps, design action and its corresponding parameters and/or analysis values, such as the coordinate of an object and/or ANE output. See an example as below:

{"Timestamp": "2017-11-14 12:51:27", "File": "EnergyPlusHome.ng3", "Add Rack": {"Type": "Rack", "Building": 2, "ID": 23, "Coordinates": [{"x": -28.863, "y": -49.8, "z": 20.799}]}}

In this study, only the design actions, e.g., "Add wall", "Edit wall", "Show heliodon", etc. are extracted for analysis. Trivial actions that do not affect the design quality, such as "Camera", "Add human", "Edit human" etc., are ignored. The participants have tried 115 different types of actions. After collecting the

[1] We conducted questionnaire and collected the demographics and other basic information about participants. Due to the length of the paper and its main focus, the statistics of the questionnaire is not reported.

[2] Our pilot study has shown that participants are able to master the operations of Energy3D for the energy-plus home design project in 30 minutes with the aid of the tutorial.

TABLE 3: FBS coding scheme for design action data

Design process-stage	Design action
Formulation	Add any component
Analysis	Analysis of annual net energy
Synthesis	Edit any component
Evaluation	Cost analysis
Reformulation 1	Remove structure
Reformulation 2	Remove solar device
Reformulation 3	Remove other components

action data, we develop a coding scheme (Table 3) based on the FBS-based design process model to transform the design action data to the design process data in support of the cluster analysis.

In FBS ontology, *Formulation* is the process to generate Function from requirement. In our design problem, with the provided design requirements, designers start to generate house functions by adding new components, e.g., wall and window. So, we define these actions as *Formulation*. In Energy3D, to increase solar energy (i.e., the expected behavior), modification of different fictional components is required. So, *Synthesis* in our context corresponds to editing actions, e.g., change height, edit wall, etc. *Analysis* indicates the process of generating behavior from structure. In Energy 3D, such a process refers to ANE analysis of a given house structure. During the design, designers evaluate the overall design quality by comparing the ANE per dollar cost of different design alternatives. Therefore, give the same ANE, the action of doing cost analysis indicates the *Evaluation* process. Finally, in Energy3D, designers recreate structure by removing old structural components. Solar panels are sometimes removed to more precisely adjust the roof space in order to put more solar panels so as to profuce more solar energy. Therefore, in our design problem, *Reformulation 1, 2 and 3* refers to removing structure, solar devices, and other miscellaneous components (e.g., roof, tree, etc.), respectively. A complete coding scheme for this study in shown in Table 3.

5 RESULTS AND DISCUSSION

5.1 Clustering Sequential Decision Making based on Markov Chain Model

To quantify designers' sequential decision-making behaviors, the first order Markov chain transition probability [9] is calculated. An entry of the matrix π_{ij} defines the probability that design process i transitions to j, which is calculated by the following equation,

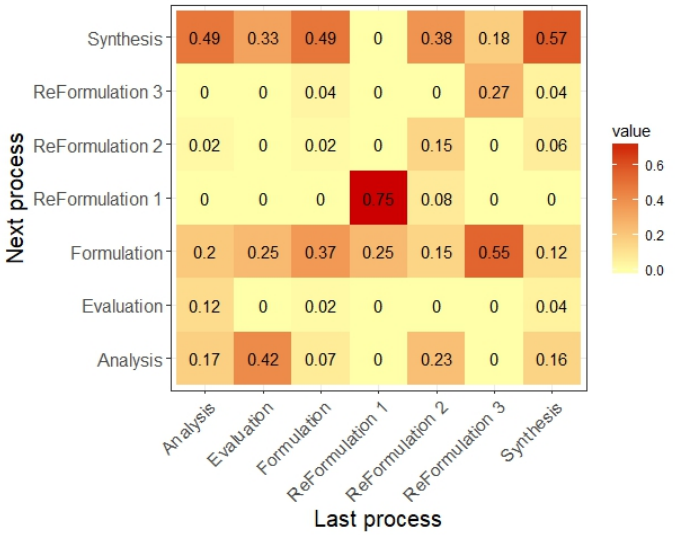

FIGURE 6: Transition matrix of the first-order Markov chain for participant C14

$$\pi_{ij} = \frac{n_{ij}}{n_i} \tag{1}$$

,where n_{ij} is the number of times design process j is followed by process i. n_i is the total counts of the process i during the entire design.

As an example, Figure 6 shows the transition probability of designer C14. It shows that the most occurred transition is *Reformulation 1* $\rightarrow$ *Reformulation 1* and the value is 0.75. This indicates that the designer C14 was involved in removing structure (wall, window) significantly more frequent than other transitions. The value zero means that the designer never used that transition in the design. For example, the value from *Synthesis* to *Reformulation 1* is zero. This indicates that after editing or changing the parameters of any structural components (such as walls), this designer would never removed those components.

Once all the 38 participants' transition probability matrices are obtained, they are converted to a 49×38 matrix that captures the sequential design process features, from which different clustering methods are applied. The optimal numbers of clusters for K-means clustering are 4, 5 and 6, which is obtained from the elbow plot technique. This means these three points correspond to the transition region where the change of the slope on the elbow plot curve is the largest. In this paper, we evaluate different clustering methods at each of the three clustering settings. Figure 7 shows the K-means clustering results with 4 groups. The clusters are indicated by four different symbols (1, 2, 3, and 4). The number of designers in each cluster is 15, 11, 10 and 2, respectively.

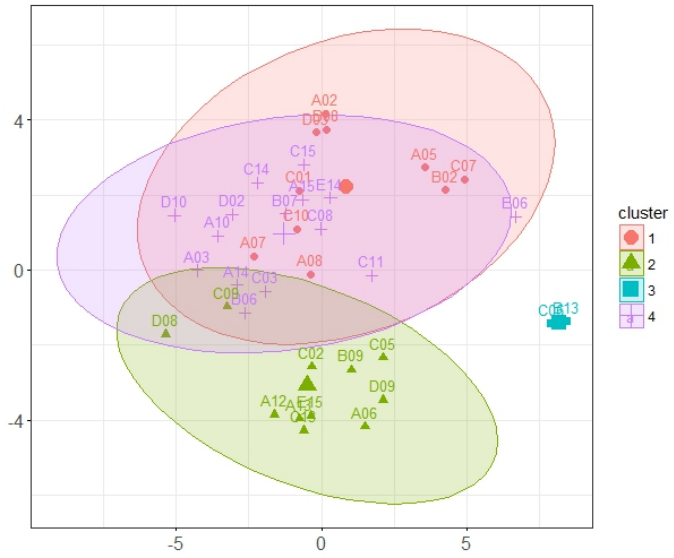

FIGURE 7: K-means clustering plot of four groups

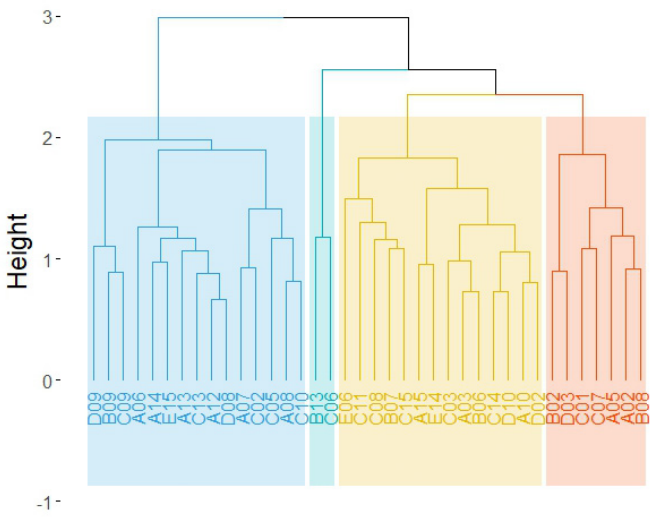

FIGURE 8: Dendrogram produced by hierarchical agglomerative algorithm

The plot shows the data points in two principle dimensions. From the figure, it is observed that designers B13 and C06 in Cluster 3 are situated far from the other clusters in the Euclidean space. It is inferred that their sequential behaviors are quite different from the other designers.

Hierarchical method clusters the designers by forming a dendrogram, as shown in Figure 8. The height of the dendrogram indicates the designers' behavioral similarity. To get 4 clusters, the dendrogram is cut at the height of 2.1. The resulting clusters contains 15, 14, 9 and 2 members, respectively. Figure 8 indicates that designer A12 and D08 meet at the lowest distance (the lowest height) on the dendrogram than any other pairs. Therefore, they share the most similarity in sequential behaviors. Like K-means-4 clustering, Hierarchical-4 (HAC-4) clustering proves the similarity between B13 and C06 as well. While in K-means-4 clustering, A10 and A14 are in the same group, but in the HAC-4 clustering they are located at two different groups. This reveals that the inconsistency among different clustering methods.

For the network-based clustering, we calculate the RSS and CS similarities between each pair of designers using the vectors obtained from the transition probability matrix. This process produces two 38×38 similarity matrices from which the RSS-based network and the CS-based network can be obtained, respectively. To obtain the desired number of clusters (i.e., 4, 5 and 6 determined by elbow plot method), we trial and error the RSS and CS values together with the modularity-maximization algorithm to determine the threshold. The results suggest that the values 1.24, 1.23 and 1.22 of RSS similarity are able to create 4, 5 and 6 clusters, respectively for RSS-based network. In the CS-based network, it is found that the values of 0.7, 0.75 and 0.77, are the appropriate threshold values to produce the desired number of clusters 4, 5, and 6.

Figure 9 shows the result of RSS-based network clustered in 4 groups indicated by different colors. The four groups consist of 14, 11, 11 and 2 members, respectively. But in this method, the clustering results are different from K-means-4 and HAC-4. For example, E06 and E14 belong to the same group in K-means-4 and HAC-4, but in RSS-4, they are in separate groups. But results from different methods do hold consistency. For example, B13

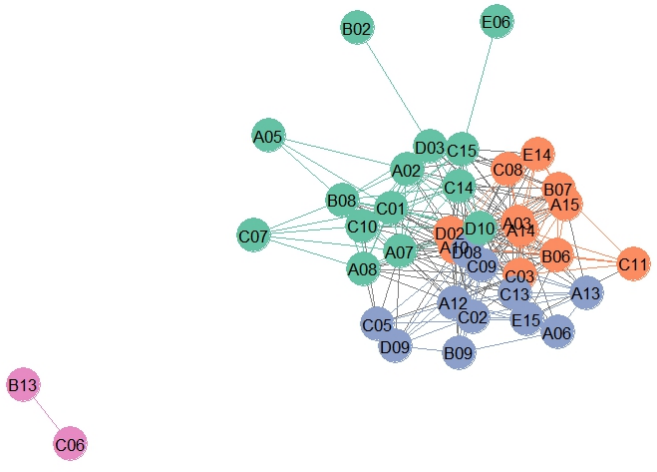

FIGURE 9: The network-based clustering using residual sum of square similarity groups the designers in four clusters

TABLE 4: Comparison of different clustering methods using variation of information. The row and column names indicate the cluster method and corresponding number of clusters.

	KM-4	KM-5	KM-6	HAC-4	HAC-5	HAC-6	RSS-4	RSS-5	RSS-6	CS-4	CS-5	CS-6
KM-4	~	~	~	0.546	0.687	0.710	0.449	1.200	0.946	0.760	0.273	0.310
KM-5	~	~	~	0.894	0.815	0.752	0.785	0.847	0.594	0.919	0.596	0.633
KM-6	~	~	~	1.250	1.170	1.107	0.985	0.737	0.886	0.862	0.936	0.955
HAC-4	0.546	0.894	1.250	~	~	~	0.978	1.469	1.130	1.250	0.696	0.733
HAC-5	0.687	0.815	1.170	~	~	~	1.120	1.389	1.050	1.392	0.820	0.857
HAC-6	0.710	0.752	1.107	~	~	~	1.142	1.334	0.836	1.414	0.757	0.794
RSS-4	0.449	0.785	0.985	0.978	1.120	1.142	~	~	~	0.953	0.821	1.331
RSS-5	1.200	0.847	0.737	1.469	1.389	1.334	~	~	~	0.647	1.419	0.673
RSS-6	0.946	0.594	0.886	1.130	1.050	0.836	~	~	~	0.684	1.419	0.710
CS-4	0.760	0.919	0.862	1.250	1.392	1.414	0.953	0.647	0.684	~	~	~
CS-5	0.273	0.596	0.936	0.696	0.820	0.757	0.821	1.419	1.419	~	~	~
CS-6	0.310	0.633	0.955	0.733	0.857	0.794	1.331	0.673	0.710	~	~	~
Efficiency	5	3	0	2	1	0	1	2	2	2	3	3

and C06 have been always grouped together in all three methods. Following the same approach of generating RSS-based network clustering, clusters can also be produced using CS-based network clustering method. CS-based clustering shows some similarities and dissimilarities as well. For example, B13 and C06 are clustered together with K-means-4, HAC-4 and RSS-4 methods, but they are separated with CS-6 method.

Since clustering results are inconstant from different clustering methods, the results need to be verified. The variation of information (VI) is used to compare each pair of clustering methods to evaluate the partial agreement between the clusters obtained from each method. The VI values are summarized in Table 4. Please note that the VI between the same clustering methods but different cluster numbers (e.g. K-means-4 vs. K-means-5) is not worth comparing, thus the corresponding VI are not available in Table 4. From Table 4, we can observe that the VI between K-means-4 clustering and CS-6 clustering is 0.31. On the other hand, the VI between HAC-5 and CS-5 clustering have is 0.82. So, K-means-4 clustering and CS-6 clustering have more overlapping cluster members than that HAC-5 and CS-5 clustering has.

By analyzing the distribution of VI (see Figure 10), the value of 0.7 (corresponding to the top 25 % quantile) is chosen as a cutoff value to filter out the clustering methods that have more consistent results. During this process, we are able to a) find the most efficient clustering method and its corresponding number of clusters, and b) find the designers that have been always clustered together and identify their sequential behavioral patterns. In Table 4, the VI values below 0.7 are highlighted in yellow color and their corresponding clustering methods are selected for further

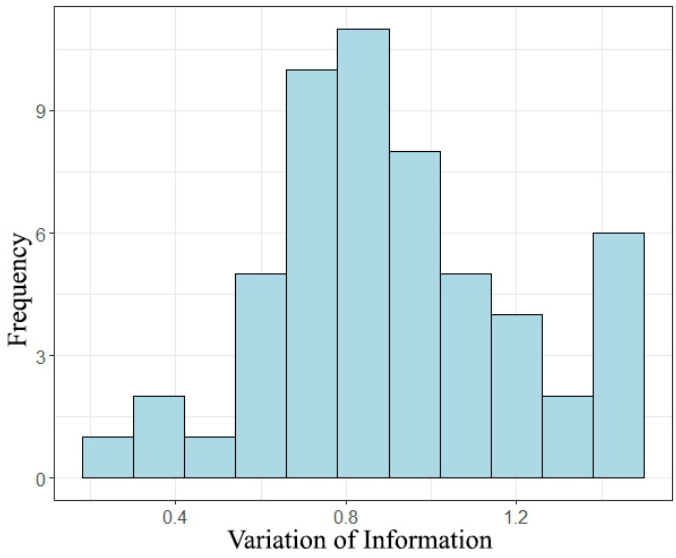

FIGURE 10: Distribution of the VI shown in Table 4

consideration. The values which are below 0.7 are considered as efficient. This can be expressed as the following way:

$$Efficiency = \sum_{i}^{k} f(VI_i) \tag{2}$$

,where $f(VI_i) = 1$ if $VI_i < 0.7$; and 0 otherwise. i= 1, 2.... k and k = 12 in this case study. It is observed that K-means-4 clustering has the largest number of times in overlapping with other clustering methods. Therefore, K-means-4 clustering is the most efficient method among all the three methods in consideration. Detailed results of K-means-4 clustering is presented earlier. By checking the occurrence of the VI values being below the threshold 0.7, we identify K-means (4, 5), HAC (4, 5), RSS-(4, 5, 6) and CS-(5, 6) for consideration to identify the designers who have been always clustered together irrespective to the methods being used. The results are shown in Table 5. Note that each row of designers are grouped together without any pre-knowledge.

With the clustering results, we revisit the first question we aim to answer in this study: *What are the most frequent sequential design behavioral patterns that most designers would follow in systems design?* By analyzing the clusters, it found that, for most of the cases, the highest transition probability for each designer in a group is similar. The sequential design behaviors that most designers follow are listed and discussed below.

- *Synthesis* $\rightarrow$ *Synthesis*
  This transition of design stages is the most frequently occurred pattern. For example, the highest transition probability of all the designers of the third group (A06, A12, A13, C13, D08, and E15) is *Synthesis* $\rightarrow$ *Synthesis*. Again, the fifth group (B11, C06) also uses this pattern very often. It indicates that the designers of these groups kept modifying the parameters of the components. The possible reason for this deign pattern is that designers are incentivized by the rewarding mechanism in the experiment, thus they tried their best to exploit the design space by sequentially changing the design parameters.
- *Reformulation* $\rightarrow$ *Formulation*
  Designers also used this pattern very frequently. We found that, the highest transition probability of the second group (A03, A15, B07, C08, D02, D10, and E14) is *Reformulation 2* $\rightarrow$ *Reformulation*. This pattern indicates that designers in this group spent significant amount of time to remove solar panels and again adding them back. It may be due to that they were trying to adjust the solar panel on the roof to a perfect condition. Again, the last group (B09, D09) followed the *Reformulation 3* $\rightarrow$ *Formulation* design pattern. Designers in this group spent most of the time to remove the existing roof or others component (excluding solar panels and structural components) and again adding it.

TABLE 5: Clustering results of design sequences irrespective of the clustering methods

A02, A05, B08, C01, C07
A03, A15, B07,C08, D02, D10,E14
A06, A12, A13, C13, D08, E15
A07, A08, C10
B06, C11
B09, D09

5.2 Clustering Design Behaviors based on the Distribution of Design Process Stages

The second question we'd like to answer is that, *If designers behave similarly in sequential design-making of time domain, would they also have similar behaviors in frequency domain?* To answer this question, we apply the same approach in Figure 1 to identify the designers who use similar number of design process stages during their designs. The only difference between this analysis and the one in the previous section is that the behavioral data used in this section in a 7×1 vector. Each element of this vector is the frequency of each design process stage. Therefore, this analysis capture similarities of designers who have similar preferences of leveraging certain design processes in systems design. Figure 11 shows the examples of the distributions of the design process stages from four designers. These results verify that our approach is able to successfully cluster the similar design behaviors together.

Table 6 shows the designers who have been always grouped together based on their design process distribution irrespective of the clustering methods. Among all the participants, it is found that *Synthesis* is the most frequently used design process stage. Out of the 10 similar behavioral groups, 9 groups follow this trend. That means, in their design processes most of the time they are involved in editing various component of the energyplus home. Such a behavior is again a reflection of the reward incentive created in the experiment. However, as shown in Figure 11, B08 and B09 do not follow this trend. Instead, the most frequent design process stage is *Formulation* which signifies that their design was much involved adding components to meet the design requirements. Participants B13 and E06 have a unique distribution of the design process stages. Instead of using all seven design processes, they are mainly involved in *Formulation*, *Synthesis*, *Analysis*, and *Evaluation* process. They almost never performed any actions related to reformulation. This indicates

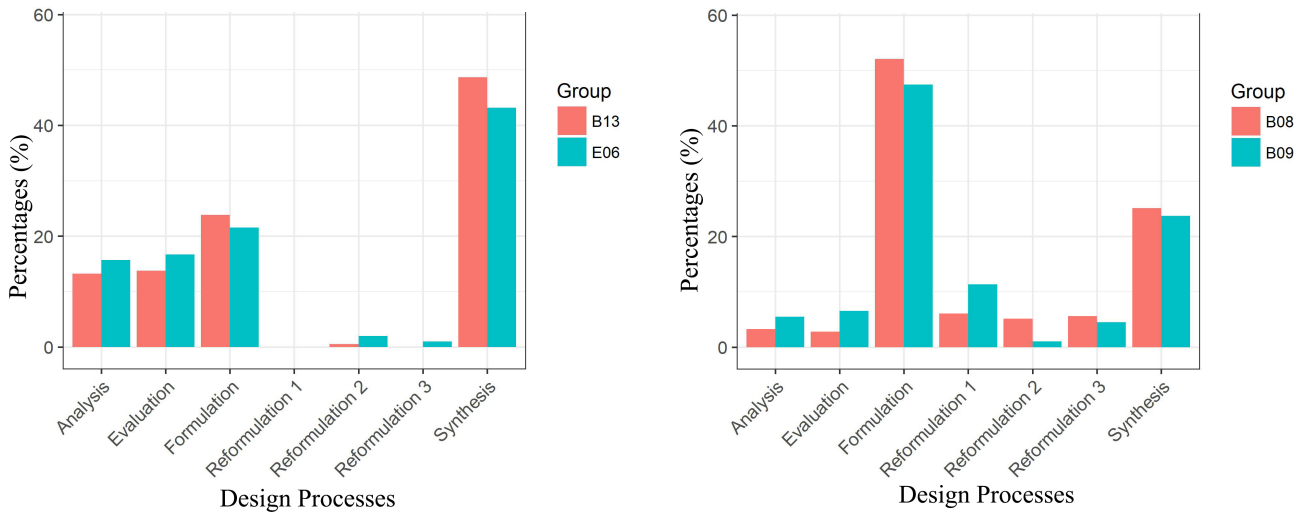

FIGURE 11: Design process stage distribution of two groups where designers in the same group show similar patterns of distribution whereas the behavioral patterns are different between groups

that different designers have different patterns and designers do have preferences in selecting certain types of design actions to explore the design space. The resulting distribution of design actions is therefore not uniform. It is observed that *Reformulation* is overall used less frequently than other process stages on average. This implies that designers are incline to improving the design quality by editing the artifacts that are already established rather than removing and restructuring the house. Some of the designers (e.g., A05, C05) performed *Analysis* almost the smae number of times as *Synthesis* and *Formulation*. This behavior indicates that they were exploring the effects of changing certain parameters because any changes made in Energy3D can be immediately assessed.

By comparing Table 5 and Table 6 it is found that only A12 and E15 grouped together in both sequential behavioral analysis and distribution analysis of design process stages. This indicates that, for most designers, even if they behave similarly in sequential design-making of time domain, they do not necessarily have similar behaviors in frequency domain.

TABLE 6: Clustering results of design process distribution irrespective of the clustering methods

A02, A14, C15, D02
B02, C01
A06, B06, B07, C02, C10, C13,D03
A10, D10
A12, E15
A05, C05
A08, C09, C11
B08, B09
A13, C14
B13,E06

6 CONCLUSIONS AND FUTURE WORK

This paper presents a framework of automatically clustering designers with similar design behaviors. Fine-grained design action data are collected using Energy3D in an non-intrusive way. Then, the first-order Markov chain is used to generate the sequential behavioral data after applying the FBS-based coding scheme. On the other hand, based on the distribution of design process stages, we analyzed the designers' behaviors quantified in frequency domain. We utilized three representative clustering methods, K-means, the hierarchical agglomerative, and the network-based clustering methods in this study. The elbow plot method indicates that 4, 5 and 6 are preferred clustering numbers. In order to verify the clustering results, variation of information method is used and we find that K-means with 4 clusters is the most efficient clustering method. Finally, by comparing the obtained clusters, designers with similar sequential behav-

ioral patterns are identified. We find that, *Synthesis* $\rightarrow$ *Synthesis* and *Reformulation* $\rightarrow$ *Formulation* are the design patterns that were followed by a large number of designers. In addition, we find that designers who used the same number of process stages do not necessarily follow the same sequence in their design.

The overall contribution of this paper is the development of a general framework that can accommodate various clustering methods for identifying design behavioral patterns. Moreover, the network-based clustering approach developed in this study provides a new way for clustering design behaviors by leveraging network community-detection algorithms. Successful identification of similar behaviors as well as their design patterns has significant benefits in discovering efficient design heuristics and guiding team-based design. For example, useful design process-stage frequencies and design patterns that lead to better design outcomes can be identified by correlating design quality with different behavioral groups. Also, in team-based design, to maximize the working efficiency, similar/dissimilar designers could be paired up to improve the communication and /or diversity within a group.

In the future work, more concrete validation study will be performed. On the one hand, the potential factors, such as designers' demographics and expertise, which result in the observed clusters will be studied. This helps further validate the correctness of the clustering results and identify the influential factors that drive the formation of clusters. On the other hand, the clustering results obtained from this study can be used in other applications, for example in the prediction of sequential design behaviors, to further demonstrate the usefulness of the cluster information. In addition, we plan to evaluate the correlation between design sequences and design quality in order to identify beneficial sequential decision strategies. Also, to understand sequential behavior more precisely, higher-order Markov chain will be applied to study the memory effects in sequential design behaviors. Finally, we are interested in exploring other possible models in addition to the Markov chain to quantify the sequential decisions so that the robustness of the proposed clustering framework can be evaluated.

7 ACKNOWLEDGEMENTS

We acknowledge the financial support from the 2017 Engineering Research and Innovation Seed Funding Program from the College of Engineering at the University of Arkansas. We also acknowledge Kaleb Porter and Luke Godfrey for their technical support and sharing the pearls of wisdom during many research meetings of this project.

REFERENCES

[1] Austin-Breneman, J., Honda, T., and Yang, M. C., 2012. "A study of student design team behaviors in complex system design". *Journal of Mechanical Design,* ***134***(12), p. 124504.

[2] Dym, C. L., Agogino, A. M., Eris, O., Frey, D. D., and Leifer, L. J., 2005. "Engineering design thinking, teaching, and learning". *Journal of Engineering Education,* ***94***(1), pp. 103–120.

[3] Brockmann, E. N., and Anthony, W. P., 1998. "The influence of tacit knowledge and collective mind on strategic planning". *Journal of Managerial issues*, pp. 204–222.

[4] Smith, R. P., and Eppinger, S. D., 1997. "A predictive model of sequential iteration in engineering design". *Management Science,* ***43***(8), pp. 1104–1120.

[5] Yukish, M. A., Miller, S. W., and Simpson, T. W., 2015. "A preliminary model of design as a sequential decision process". *Procedia Computer Science,* ***44***, pp. 174–183.

[6] Sha, Z., Kannan, K. N., and Panchal, J. H., 2015. "Behavioral experimentation and game theory in engineering systems design". *Journal of Mechanical Design,* ***137***(5), p. 051405.

[7] Yu, R., Gero, J. S., Ikeda, Y., Herr, C., Holzer, D., Kaijima, S., Kim, M., and Schnabel, A., 2015. "An empirical foundation for design patterns in parametric design". In 20th International Conference of the Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Daegu, South Korea, May, Citeseer, pp. 20–23.

[8] Kan, J. W. T., and Gero, J. S., 2011. "Comparing designing across different domains: An exploratory case study". In DS 68-2: Proceedings of the 18th International Conference on Engineering Design (ICED 11), Impacting Society through Engineering Design, Vol. 2: Design Theory and Research Methodology, Lyngby/Copenhagen, Denmark, 15.-19.08. 2011.

[9] McComb, C., Cagan, J., and Kotovsky, K., 2017. "Capturing human sequence-learning abilities in configuration design tasks through markov chains". *Journal of Mechanical Design,* ***139***(9), p. 091101.

[10] Blessing, L. T., 1995. "Comparison of design models proposed in prescriptive literature".

[11] Asimow, M., 1962. *Introduction to design*. Prentice-Hall.

[12] Darke, J., 1979. "The primary generator and the design process". *Design studies,* ***1***(1), pp. 36–44.

[13] March, L., 1984. The logic of design* in cross, n.(ed.): Developments in design methodology.

[14] Pahl, G., and Beitz, W., 2013. *Engineering design: a systematic approach*. Springer Science & Business Media.

[15] Council, D., 2005. "The 'double diamond'design process model". *Design Council*.

[16] Howard, T. J., Culley, S. J., and Dekoninck, E., 2008. "Describing the creative design process by the integration of engineering design and cognitive psychology literature". *Design studies,* ***29***(2), pp. 160–180.

[17] Bunge, M., 1977. Treatise on basic philosophy: Ontology

i: The furniture of the world. dordrecht, holland: D.
[18] Gero, J. S., 1990. "Design prototypes: a knowledge representation schema for design". *AI magazine,* ***11***(4), p. 26.
[19] Umeda, Y., Tomiyama, T., and Yoshikawa, H., 1995. "Fbs modeling: modeling scheme of function for conceptual design". In Proc. of the 9th Int. Workshop on Qualitative Reasoning, pp. 271–8.
[20] Deng, Y.-M., Tor, S. B., and Britton, G., 1999. "A computerized design environment for functional modeling of mechanical products". In Proceedings of the fifth ACM symposium on Solid modeling and applications, ACM, pp. 1–12.
[21] Kan, J. W., and Gero, J. S., 2009. "Using the fbs ontology to capture semantic design information in design protocol studies". In About: Designing. Analysing Design Meetings, CRC Press, pp. 213–229.
[22] Kodinariya, T. M., and Makwana, P. R., 2013. "Review on determining number of cluster in k-means clustering". *International Journal,* ***1***(6), pp. 90–95.
[23] Stroock, D. W., 2013. *An introduction to Markov processes*, Vol. 230. Springer Science & Business Media.
[24] Karlin, S., 2014. *A first course in stochastic processes*. Academic press.
[25] Tan, P.-N., Steinbach, M., and Kumar, V., 2013. "Data mining cluster analysis: basic concepts and algorithms". *Introduction to data mining*.
[26] Äyrämö, S., and Kärkkäinen, T., 2006. "Introduction to partitioning-based clustering methods with a robust example". *Reports of the Department of Mathematical Information Technology. Series C, Software engineering and computational intelligence 1/2006*.
[27] Chaoji, V., Al Hasan, M., Salem, S., and Zaki, M. J., 2008. "Sparcl: Efficient and effective shape-based clustering". In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, IEEE, pp. 93–102.
[28] Berkhin, P., 2006. "A survey of clustering data mining techniques". In *Grouping multidimensional data*. Springer, pp. 25–71.
[29] Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A., 2011. "Density-based clustering". *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* ***1***(3), pp. 231–240.
[30] Schaeffer, S. E., 2007. "Graph clustering". *Computer science review,* ***1***(1), pp. 27–64.
[31] Singh, V. K., Tiwari, N., and Garg, S., 2011. "Document clustering using k-means, heuristic k-means and fuzzy c-means". In Computational Intelligence and Communication Networks (CICN), 2011 International Conference on, IEEE, pp. 297–301.
[32] James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013. *An introduction to statistical learning*, Vol. 112. Springer.
[33] Steinbach, M., Karypis, G., Kumar, V., et al., 2000. "A comparison of document clustering techniques". In KDD workshop on text mining, Vol. 400, Boston, pp. 525–526.
[34] Chen, M., Kuzmin, K., and Szymanski, B. K., 2014. "Community detection via maximization of modularity and its variants". *IEEE Transactions on Computational Social Systems,* ***1***(1), pp. 46–65.
[35] Brandes, U., Delling, D., Gaertler, M., Gorke, R., Hoefer, M., Nikoloski, Z., and Wagner, D., 2008. "On modularity clustering". *IEEE transactions on knowledge and data engineering,* ***20***(2), pp. 172–188.
[36] Wagner, S., and Wagner, D., 2007. *Comparing clusterings: an overview*. Universität Karlsruhe, Fakultät für Informatik Karlsruhe.
[37] Egan, P., and Cagan, J., 2016. "Human and computational approaches for design problem-solving". In *Experimental Design Research*. Springer, pp. 187–205.
[38] Xie, C., Zhang, Z., Nourian, S., Pallant, A., and Bailey, S., 2014. "On the instructional sensitivity of cad logs". *International Journal of Engineering Education,* ***30***(4), pp. 760–778.
[39] Xie, C., Zhang, Z., Nourian, S., Pallant, A., and Hazzard, E., 2014. "A time series analysis method for assessing engineering design processes using a cad tool". *International Journal of Engineering Education,* ***30***(1), pp. 218–230.