E-Companion:

On Styles in Product Design: An Analysis of US Design Patents

PART A: FORMALIZING THE DEFINITION OF STYLES

A.1 Styles as categories of designs of similar form

Our task involves categorizing a large set of designs into categories of styles on the basis of their form similarities. Formalizing our definition of styles serves two purposes. First, it concretizes the theoretical development, and thus is instrumental in confirming the conceptual integrity of our definition. Second, the formalization proposed here serves as a theoretical anchor for validating the outcome of any style-identifying algorithm.

First, we make the following assumptions about the similarity in concept between designs:

Assumption 1: Given two arbitrary designs D_a, D_b, we can measure the similarity in form between them $(D_a, D_b) \rightarrow R^+$. Further, we assume this similarity is deterministic.

Suppose we have a set of points $\{D_1, D_2, \dots, D_n\}$, a categorization Ω is a partition of S into clusters $\{C_1, C_2, \dots, C_k\}$ such that the following properties obtain: (i) mutual exclusivity $C_i \cap C_j = \emptyset, \forall\, i$, and (ii) completeness $\cup_{i=1}^{k} C_i = S$. Implicit in this setup are the following assumptions:

Assumption 2: Styles are represented by discrete categories.

Assumption 3: There exist a categorization Ω_1 such that:

(i) All clusters in Ω_1, $C_i \in \Omega_1, \forall\, i$, are considered styles.

(ii) Any pairs of clusters taken from Ω_1 must be stylistically distinct, that is, the set of designs formed by merging designs from the pair would not be considered a style.

Our goal is thus to identify Ω_1. Note that Assumption 1 follows from cognitive theory that there is a single holistic similarity underlying object comparisons. Assumptions 2 and 3 simplifies the world of designs by considering a discrete categorization, essentially capturing the most important phenomena of 'main styles'. Thus, we ignore variations within styles (the nested sub-styles), and the relationships across styles.

A.2 Cluster Structure

Because clustering methods need to be tailored to the data at hand, the above definition of styles is insufficient to identify an appropriate clustering method. What is additionally required is some knowledge of how styles are structured. Based on the two features of our similarity graph—(i) the similarity graph exhibits more clustering than a random graph, and (ii) the distribution of node degrees is highly skewed—we assume that our data has a similar structure as the theoretical network suggested by Ravasz and Barabási (2003), which has the above two properties and a clear nested cluster structure.

The model is represented in Figure A-1. In the model, all edges have a weight of 1. The iterative construction leading to a hierarchical network starts from a fully connected cluster of five nodes shown in (a). Four identical replicas of this cluster are created, connecting the peripheral nodes of each cluster to the central node of the original cluster, obtaining a network of N=25 nodes (b). In the next step, four replicas of the obtained cluster is created, and connecting the peripheral nodes again, as shown in (c), to the central node of the original module, obtaining a N=125-node network. This process can be continued indefinitely creating ever larger networks.

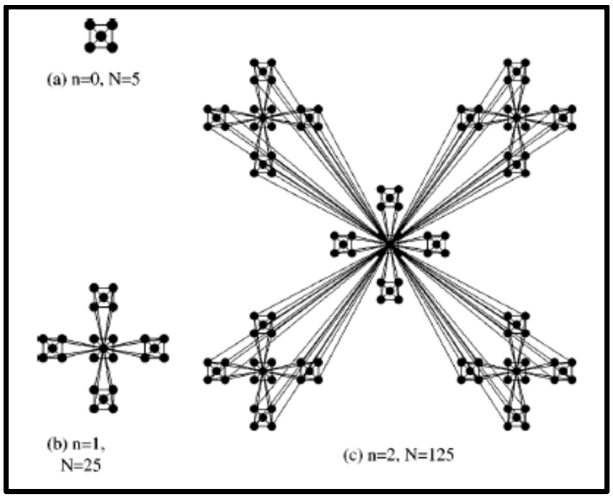


Figure A-1: The iterative construction leading to a hierarchical network of the model by Ravasz and Barabási

A.3 Conductance-based Clustering Method

For exposition, suppose we term a configuration represented by (a) as a cluster of order 0 (C_0), so that the configuration represented by (b) is a cluster of order 1 (C_1), and (c) is a cluster of order 2 (C_2), and so on. By construction, the model connects five clusters of order $C_N, N \geq 0$ to form a larger cluster of order C_{N+1} at each iteration. Suppose we have data that is represented by C_N, and we want to recover all proper categorizations nested within C_N (i.e. categorizations consisting of 5 clusters of C_{N-1}, 25 clusters of C_{N-2}, and so on, denoted as $\Omega_j, j \in \{0,1,\dots,N-1\}$, where Ω_j contains only clusters of order j and nothing else), we show in this section that the ideal clustering method employed is an iteratively divisive algorithm that optimizes *conductance*.

Given a graph G, its conductance $\Phi(G)$ (Chung, 1997) is a measure of its partition-ability. Specifically, G is more partition-able if there exist two subgraphs $g^*, \bar{g}^* \subset G$, $g^* \cup \bar{g}^* = G$ such that there is (i) low similarity across g^* and $\bar{g}^*$ (measurable by the cut function, which is the sum of similarity of edges connecting between g^* and $\bar{g}^*$), and (ii) high similarity within g^*, $\bar{g}^*$ (measurable by the volume function, which is the sum of degrees of those nodes within $g, \bar{g}^*$). $\Phi(G)$ is obtained via optimizing over all possible subgraphs $g, \bar{g}$ such that the ratio of cut over the smaller volume is minimized (see Equation A1).

$$\Phi(G) = \frac{\text{cut}(g^*, \bar{g}^*)}{\min(\text{volume}(g^*), \text{volume}(\bar{g}^*))} = \min_{g \subset G} \frac{\text{cut}(g, \bar{g})}{\min(\text{volume}(g), \text{volume}(\bar{g}))} \quad \text{(A1)}$$

The following theorem states our main result:

Theorem 1: Given a cluster $C_N, N \geq 0$. An iterative divisive algorithm that does the following within its select, partition, and evaluate subroutines will obtain as its solution set $\Omega_j, j \in \{0,1,\dots,N-1\}$, and nothing else.

1) *Select* the most partition-able cluster (i.e. one with the lowest Φ, denote its conductance as Φ_C).
2) *Partition* it into $g^*, \bar{g}^*$ (the solution obtained when calculating Φ_C).

3) *Evaluate* the value of $\Phi_N - \Phi_C$, where Φ_N is the conductance of the most partition-able cluster in the resulting set of clusters. Admit the current categorization as a solution if $\Phi_N - \Phi_C > 0$.

Intuitively, **Theorem 1** says that an algorithm should, at each step, select the most partition-able cluster for partitioning, then partition it such that there is low similarity across clusters and high similarity within, and finally, take note of sharp increases in conductance as the algorithm progresses. Note that the last step is known amongst practitioners as the eigengap heuristics (von Luxburg, 2007).

A.4 Sketch Proof to Theorem 1

The proof has three parts, we first show that the algorithm would partition a cluster C_N by removing periphery subclusters. We then show, given a set of clusters with different orders, the algorithm would select the one with the highest order to partition. Both of these steps ensure that the clustering tree would contain all Ω_js. Finally, we show that the evaluation subroutine would pick out only Ω_js.

Partitioning clusters: We first show how the algorithm would partition a cluster $C_N, N > 0$. To facilitate exposition, denote $C_N \backslash i, i \in \{1,2,3,4\}$ as a network represented by C_N after removal of i periphery subclusters (note: $C_N \backslash 4 = C_{N-1}$, i.e. C_N becomes C_{N-1} after removal of all 4 peripheral subclusters). Also, define the function $\phi(g, \bar{g})$:

$$\phi(g, \bar{g}) = \frac{\text{cut}\,(g,\bar{g})}{\min(\text{volume}(g),\text{volume}(\bar{g}))} \tag{A2}$$

Lemma 1: $\Phi(C_N) = \frac{4^N}{\text{volume}(C_{N-1})} = \phi(C_N \backslash 1, C_{N-1}), N \geq 1$

This lemma shows that the optimal way to partition a cluster C_N is via removal of one periphery subcluster (C_{N-1}) from C_N, producing subclusters $\{C_{N-1}, C_N \backslash 1\}$.

We extend **Lemma 1** to examine how the algorithm would partition an *incomplete* cluster, i.e. C_N with a few periphery subclusters removed (note that the important cases are where C_N has up to three periphery subclusters removed, because having all four removed leaves clusters C_{N-1}).

Lemma 1A: $\Phi(C_N \backslash i) = \Phi(C_N) = \phi(C_N(i+1), C_{N-1}), i \in \{1,2,3\}$.

Lemma 1A states that, given a cluster with a few peripheral subclusters removed, the algorithm would continue partitioning via removal of additional peripheral subclusters (the proof of this lemma follows closely in structure to the previous one, and is omitted).

Given **Lemmas 1** and **1A**, it follows that given a cluster C_N, the algorithm will iteratively partition it until it obtains the 5 subclusters C_{N-1}, and at any point in its iteration will not destroy the cluster structure.

Selecting clusters: A clear corollary of the previous lemmas is that at any point of time in the algorithm's iterations, the algorithm can have a mix of clusters (some possibly with peripheral subclusters removed) of different orders, but nothing else, to select from. The following lemma establishes the order at which the algorithm would select clusters.

Lemma 2: $\Phi(C_N) < \Phi(C_{N-1})$, for all $N \geq 1$.

Given Lemma 2, the algorithm would thus consider a cluster with a larger N to partition first, and this is the desired behavior. Additionally, because of Lemma 1A, an incomplete cluster of order N is treated similarly as a complete cluster of order N. So the algorithm will not proceed to select any cluster of order $N-1$ before all complete and incomplete clusters of order N are completely divided into clusters of order $N-1$. The proof of this lemma is also straightforward, given the formula in Lemma 1, and is thus omitted.

Evaluating categorizations: As a consequence of the above lemmas, it is thus clear that an iterative algorithm based on conductance would contain, within its iterations, all $\Omega_j, j \in \{0,1,\dots,N-1\}$. What is lastly required is a way to pick out those Ω_js, i.e. the categorizations where we have a set of homogeneous clusters with no incomplete clusters. Lemma 3 states this formally.

Lemma 3: Suppose we have an algorithm that has the behavior as described in Theorem 1. Execute the algorithm until we have reached a categorization where all clusters are formed by C_0. Denote the categorizations evaluated as potential solutions as $\{o_1,\dots,o_n\}$—where o_1 is the first identified

categorization, and o_n represents the final categorization Ω_0 (all clusters are C_0). Then, $\{\Omega_0, \dots, \Omega_{N-1}\} = \{o_1, \dots, o_n\}$.

The lemma states that tracking jumps in conductance as the algorithm progresses is a way to identify structural breaks, i.e. regions where categorizations contain clusters of similar levels of heterogeneity. The proof of this lemma follows directly from lemmas 1, 1A, and 2, and is thus omitted.

A.5 Proof to Lemma 1

We show here a formal proof to Lemma 1.

First, we make two claims with regards to the nature of solutions generated by a conductance-based method. Given a connected graph G and a proper subset of nodes $g^* \subset G$, such that $\text{volume}(g^*) \leq \text{volume}(G \backslash g^*)$. Let $\{g^*, G \backslash g^*\}$ be the optimal partition—hence, $\Phi(G) = \phi(g^*, G \backslash g^*)$.

Claim 1: g^* is either (i) connected, or (ii) each component $\{g_1^*, \dots, g_k^*\}$ in g^* must satisfy $\frac{\text{cut}\,(g_1^*, G \backslash g_1^*)}{\text{volume}\,(g_1^*)} = \frac{\text{cut}\,(g_2^*, G \backslash g_2^*)}{\text{volume}\,(g_2^*)} = \cdots = \frac{\text{cut}\,(g_k^*, G \backslash g_k^*)}{\text{volume}\,(g_k^*)}$.

Proof: The proof goes by way of contradiction. Suppose, on the contrary, g^* is disconnected with components $\{g_1^*, \dots, g_k^*\}$, and, without loss of generality, ordered in increasing cut over volume ratio, that is, $\frac{\text{cut}\,(g_1^*, G \backslash g_1^*)}{\text{volume}\,(g_1^*)} \leq \cdots \leq \frac{\text{cut}\,(g_k^*, G \backslash g_k^*)}{\text{volume}\,(g_k^*)}$, and where at least one of the inequality is strict. Then,

$$\begin{aligned} \Phi(G) &= \frac{\text{cut}(g_1^*, G \backslash g^*) + \dots + \text{cut}(g_k^*, G \backslash g^*)}{\text{volume}(g_1^*) + \dots + \text{volume}(g_k^*)} \\ &= \sum_{j=1}^{k} \frac{\text{volume}(g_j^*)}{\sum_{i=1}^{k} \text{volume}(g_i^*)} \left[\frac{\text{cut}(g_j^*, G \backslash g^*)}{\text{volume}(g_j^*)} \right] \\ &> \frac{\text{cut}(g_1^*, G \backslash g^*)}{\text{volume}(g_1^*)} = \frac{\text{cut}(g_1^*, G \backslash g_1^*)}{\text{volume}(g_1^*)} = \phi(g_1^*, G \backslash g_1^*) \end{aligned} \tag{A3}$$

Where the first equality follows from the definition of conductance, the second equality shows that conductance is a weighted sum of cut over volume ratio of the components of the smaller side. The inequality follows from the consideration that g_1^* has the smallest cut over volume ratio. The third equality follows from the disconnectedness of components. Finally, since $\text{volume}(g_1^*) < \text{volume}(G \backslash g_1^*)$,

it is the smaller side and we have found a partition $\{g_1^*, G\backslash g_1^*\}$ produces a lower ϕ than the optimal partition of $\{g^*, G\backslash g^*\}$ (last equality), which is a contradiction. ■

Claim 2: $G\backslash g^*$ is connected.

Proof: The proof similarly goes by contradiction.

Consider first the case where the bigger side is strictly bigger. So, $\text{volume}(G\backslash g^*) > \text{volume}(g^*)$. Suppose, on the contrary, $G\backslash g^*$ is disconnected with components $\{\bar{g}_1^*, \dots, \bar{g}_k^*\}$, and, without loss of generality, ordered in increasing cut over volume ratio, that is, $\frac{\text{cut}\,(g_1^*, G\backslash g_1^*)}{\text{volume}\,(g_1^*)} \leq \cdots \leq \frac{\text{cut}\,(g_k^*, G\backslash g_k^*)}{\text{volume}\,(g_k^*)}$. Note first the inequality: $\Phi(G) = \frac{\text{cut}(g^*, \bar{g}_1^*) + \dots + \text{cut}(g^*, \bar{g}_k^*)}{\text{volume}(g^*)} > \frac{\text{cut}(g^*, \bar{g}_1^*)}{\text{volume}(g^*)}$. Note also:

$$\begin{aligned} \Phi(G) &> \frac{\text{cut}(g^*, \bar{g}_1^*) + \dots + \text{cut}(g^*, \bar{g}_k^*)}{\text{volume}(\bar{g}^*)} \\ &= \sum_{j=1}^{k} \frac{\text{volume}(\bar{g}_j^*)}{\sum_{i=1}^{k} \text{volume}(\bar{g}_i^*)} \left[\frac{\text{cut}(g^*, \bar{g}_j^*)}{\text{volume}(\bar{g}_j^*)}\right] \\ &\geq \frac{\text{cut}(g^*, \bar{g}_1^*)}{\text{volume}(\bar{g}_1^*)} \end{aligned} \tag{A4}$$

In combination, we have shown that $\Phi(G) > \frac{\text{cut}(g^*, \bar{g}_1^*)}{\text{volume}(g^*)}$ and $\Phi(G) > \frac{\text{cut}(g^*, \bar{g}_1^*)}{\text{volume}(\bar{g}_1^*)}$. Because $\bar{g}_1^*$ is disconnected from $\{\bar{g}_2^*, \dots, \bar{g}_k^*\}$, this implies $\Phi(G) > \phi(\bar{g}_1^*, G\backslash \bar{g}_1^*)$. A contradiction.

Consider next the case where $\text{volume}(G\backslash g^*) = \text{volume}(g^*)$. Here, we want to show that at least one side of the partition is connected. Assume, on the contrary, that both g^* and $G\backslash g^*$ have multiple components, i.e. $g^* = \{g_1^*, \dots, g_k^*\}$ and $G\backslash g^* = \{\bar{g}_1^*, \dots, \bar{g}_h^*\}$. A consequence of Lemma 1 is that all the components must have the same cut to volume ratio. Because G is connected, there must exist at least two components across the partition (say g_i^* and $\bar{g}_j^*$) that are connected. Consider the partition generated by putting g_i^* and $\bar{g}_j^*$ on one side, and all the other components on the other. Note that $\phi\left(g_i^* \cup \bar{g}_j^*, G\backslash \{g_i^* \cup \bar{g}_j^*\}\right) < \Phi(G)$, because those connections between g_i^* and $\bar{g}_j^*$ now do not contribute to the cut. Here we have found a better partition than the optimal. A contradiction. ■

Overall, given a connected graph G, a conductance-based algorithm would partition it such that the bigger side is connected, and the smaller side is either connected or composed of multiple components with the same cut to volume ratio. In the case where the optimal solution is an equal-volume partition, then at least one side is connected. We leverage on these two claims and a unique feature of the graph—i.e. the global center vertex of a cluster C_N, denoted V_N^C, is the only vertex that connects the periphery subclusters, denoted as $C_N^i, i \in \{1,2,3,4\}$ to the central subcluster, denoted C_N^C.

Corollary 1: A partition $\{A^*, C_N \setminus A^*\}$, $\text{volume}(A^*) \leq \text{volume}(C_N \setminus A^*)$ that solves for $\Phi(C_N)$ must be such that $V_N^C \in C_N \setminus A^*$, and $C_N \setminus A^*$ is connected.

That V_N^C should be on the bigger, connected side of the cut is straightforward. Consider the case where $C_N^* \setminus A^*$ has a strictly larger volume. In this case, if V_N^C is on the smaller side, the biggest volume that could be achieved on the bigger side while ensuring that the bigger side is connected is to take an entire periphery subcluster. The volume on the bigger side that can be achieved by doing this is only about one-fifth of the total volume of C_N, which will not allow the bigger side to have a volume larger than the volume of the smaller side. Alternatively, if the solution is an equal-volume partition, then it is clear that the side *without* V_N^C must be disconnected, meaning the side with V_N^C must be connected.

Our proof of **Lemma 1** goes by way of induction. First, we prove that the solution is true for C_1. Next, assuming the solution is true for $C_1, \dots, C_{N-1}$, we show that this implies that the solution is true for C_N. Then, by induction, the solution is true for all N.

Solution is true for C_1: Our goal is to first show that a solution for $\Phi(C_1)$ is given by $\phi\left(C_0^i, C_1 \setminus C_0^i\right), \forall\, i$ (removal of any one periphery subcluster) or $\phi\left[C_0^i \cup C_0^j, C_1 \setminus \left(C_0^i \cup C_0^j\right)\right], \forall\, i \neq j$ (removal of two periphery subclusters). Then, $\Phi(C_1)$ can easily be calculated from the solution.

Consider the case where some vertices v of C_0^1 is such that $v \in A^*$. Then, it is easy to observe that $\min[\text{cut}(v, C_1 \setminus A^*)] = 4$. First, note that all the periphery subclusters are only connected via V^C, meaning how C_0^2, C_0^3, C_0^4 are partitioned do not influence $\text{cut}(v, C_1 \setminus A^*)$, since they are disconnected from

v if $V^C \in C_1 \backslash A^*$. Second, since C_0^1 is fully connected, every vertex has at least four edges to other vertices in C_0^1, and thus, any partition of C_0^1 into two non-empty sets must have at least a cut of value 4. Suppose C_0^1 is entirely in A^*, then the cut value is also 4, since, by construction, C_0^1 is connected to V_1^C with four edges.

So, $\min \frac{\text{cut}(v, C_1 \backslash A^*)}{\text{volume}(v)} = \frac{cut(C_0^1, C_1 \backslash A^*)}{\text{volume}(C_0^1)} = \frac{4}{\text{volume}(C_0^1)}$. Thus, supposing only vertices of C_0^1 is in A^*, the solution for conductance must be given by $\{C_0^1, C_1 \backslash C_0^1\}$.

If additionally, we have vertices of other sub-clusters in A^*, then, they are always disconnected to C_0^1, since they are all connected only via V^C, and $V^C \in V \backslash A^*$. By Lemma 1, if vertices of other subclusters are in A^*, they must (i) have the same cut to volume ratio, and (ii) have the same overall conductance. The implication is then the entire subcluster must be in A^*.

Note that the central sub-cluster C_0^C also contains four more vertices other than V_0^C. Here, it is easy to check that $C_0^C \subset V \backslash A^*$, since it is disconnected from C_0^1 and it cannot achieve a cut to volume ratio of $\frac{4}{\text{volume}(C_0^1)}$.

Note that two sub-clusters in A^* is the maximum at which $\text{volume}(A^*) \leq \text{volume}(V \backslash A^*)$. Thus, the solution must be given by either $\phi(C_0^i, C_1 \backslash C_0^i), \forall\, i$ or $\phi[C_0^i \cup C_0^j, C_1 \backslash (C_0^i \cup C_0^j)], \forall\, i \neq j$.

Solution is true for general C_N: Suppose now the solution is generally true for $C_1, \ldots, C_{N-1}$. Consider then C_N.

For reasons of exposition, let $V_i^P, i \in \{1,2,3,4\}$ denote the center vertex of the four periphery subclusters of C_N. Our solution approach is similar to that of C_1, where we show that if vertices of one periphery subcluster C_{N-1}^i is in A^*, then it must be that C_{N-1}^i is in A^*. However, we consider this in two cases: (i) $V_i^P \in A^*$ and (ii) $V_i^P \in V \backslash A^*$.

Consider case (i). Let v be the set of vertices belonging to C_{N-1}^i to be in A^*. Assume that $V_i^P \in v$. We can observe that $\min[cut(v, C_{N-1}^i \backslash A^*)] = 4^N$ (we can see this via noting that all paths from V^C to

V_i^P intermediates all the periphery nodes in C_{N-1}^i, of which there are 4^N of them). Further, having the entire subcluster C_{N-1}^i in A^* has a cut value of 4^N. Thus, $\min \frac{\text{cut}(v, C_N \setminus A^*)}{\text{volume}(v)} = \frac{4^N}{\text{volume}(C_{N-1}^i)}$.

Consider then case (ii). Again, let v be the set of vertices belonging to C_{N-1}^i to be in A^*. However, here we assume that $V_i^P \notin v$. Notice that this implies that the elements of v in A^* that belongs to the five nested clusters of order $N-2$ (if they are in A^*) must be now disconnected. Consider then a subset of vertices u in $u \subseteq C_{N-2}^i \subseteq v$. The minimum cut over volume ratio of this set of vertices must be such that $\min \frac{\text{cut}(u, C_N \setminus A^*)}{\text{volume}(u)} \geq \frac{\text{cut}(C_{N-2}, C_{N-1} \setminus C_{N-2})}{\text{volume}(C_{N-2})}$. This can be established by observing that the relationship between C_{N-2}^i to the global center V^C is symmetric to that of the peripheral center V_i^P. Now that both V_i^P and V^C are in $C_N \setminus A^*$, this "sub-sub-cluster" has the same number of internal connections, but now double the number of edges crossing from the peripheral vertices to $C_N \setminus A^*$. Since we assume that the conductance solution is true for $N-2$, and since this change cannot lead to a lower cut to volume ratio, we have the inequality.

Further, since volume increases by at least five times for each increase in order, we have:

$$\frac{\text{cut}(C_{N-2}, C_{N-1} \setminus C_{N-2})}{\text{volume}(C_{N-2})} = \frac{4^{N-1}}{\text{volume}(C_{N-2}^i)} > \frac{4^N}{\text{volume}(C_{N-1}^i)} \quad \text{(A5)}$$

Thus, case (i) has a lower conductance value than case (ii), in which case it is the one that minimizes cut over volume ratio.

Note that, connected to V^C are also an array of sub-clusters of smaller orders (directly nested to C_{N-1}^C). Since, by our assumption their cut to volume ratio is maximized by having the entire sub-cluster in A^*, their cut to volume ratio are given by $\frac{4^j}{\text{volume}(C_{j-1})}, j < N$. This is always greater than $\frac{4^N}{\text{volume}(C_{N-1}^i)}$, thus, they cannot be part of the solution.

Note that two sub-clusters in A^* is the maximum at which $\text{volume}(A^*) \leq \text{volume}(V \setminus A^*)$, thus, the solution must be given by either $\phi\left(C_{N-1}^i, C_N \setminus C_{N-1}^i\right), \forall\, i$ or $\phi\left[C_{N-1}^i \cup C_{N-1}^j, C_N \setminus \left(C_{N-1}^i \cup C_{N-1}^j\right)\right], \forall\, i \neq j$.

Finally, since we show the proposition is true for C_1, and if the proposition is true for $C_1, \ldots, C_{N-1}$, it is also true for C_N, by induction, the proposition is true. ■

PART B: THE ALGORITHM

B.1 Construction of the similarity matrix

As input, we have a reference matrix R in which each element R_{ij} (i.e., row i and column j) is 1 if patent i referenced patent j and is 0 otherwise. Note that R is a lower triangular matrix. Denote by D a diagonal matrix in which each element along its diagonal represents the *outdegree* (i.e., the total number of outgoing references, or the sum of the row) of each patent. For expositional convenience we also denote by $\mathbf{1}$ a square matrix consisting of 1s everywhere.

The similarity matrix S is constructed as follows.

1. Improve granularity by weighing each reference by the number of outgoing references: RD^{-1}.
2. Improve symmetry by adding the matrix transposed: $RD^{-1} + (RD^{-1})^{\mathrm{T}}$.
3. Improve completeness by calculating the proportion of overlap in references.
    a. Intersection of references: $I = RR^{\mathrm{T}}$.
    b. Union of references: $U = \mathbf{1} - (\mathbf{1} - R)(\mathbf{1} - R)^{\mathrm{T}}$.
    c. Proportion of overlap: $I \backslash U$ where the operator "\" is the element-wise division of two matrices; when the value is undefined (i.e., if neither patent i nor patent j has any outgoing references), the corresponding element's default value is 0.

The similarity matrix is now given by $S = RD^{-1} + (RD^{-1})^{T} + I \backslash U$. Note that all elements of S fall within the range [0, 1].

B.2 The Select and Partition steps

The material in this section is adapted from Ng, Jordan, and Weiss (2002).

Input: Similarity matrix $S \in R^{n \times n}$.

1. Compute the normalized graph Laplacian matrix $L = I - D^{\frac{1}{2}} S D^{\frac{1}{2}}$ where *I* is the identity matrix and D is a diagonal matrix with each entry the degree of the vertex.
2. Compute the two smallest eigenvalues $\{e_1, e_2\}$ and their associated eigenvectors $\{u_1, u_2\}$
3. Let $U \in R^{n \times 2}$ be the matrix containing $\{u_1, u_2\}$ as columns.
4. Normalize the rows of U to unit lengths (i.e., length = 1).
5. Treat every row as a point in space, and use K-means (two groups) to group the *n* data points.

The algorithm approximately solves for the subgraphs g and $\bar{g}$ such that conductance is minimized; the interested reader can refer to Chung (1997) or von Luxburg (2007) for a proof. Intuitively, the approximation is due to relaxing the constraint that a design belongs strictly to either g^* or $\bar{g}^*$ (the optimal partition that minimizes conductance); that is, the solution allows a design to be (say) 80% from g^* and 20% from $\bar{g}^*$. Because the rows of $\{u_1, u_2\}$ is a rotational transformation of these membership percentages, we can use a simpler (K-means) clustering algorithm to group those designs that are close in membership. The output is then used to partition the cluster. Finally, note that e_2 is an approximation of conductance, which is used in the select step as a measure of cluster heterogeneity.

B.3 Identifying regime changes in the Evaluate step

We calculate changes in the level of conductance—known as the *eigengap heuristic* (von Luxburg 2007)—to highlight categorizations where clusters appear to be at similar levels of homogeneity in the Evaluate step. Intuitively, this method tracks the conductance of the most heterogeneous cluster in each iteration, and locates sharp jumps in conductance as the algorithm progresses. A jump indicates that it will

be significantly harder to partition the next cluster than the current one; hence a decision to continue partitioning would produce clusters at a finer level.

In order to find good candidate solutions for the main styles (the first categorization that humans consider styles) we first need to define a range of iterations from which to search since the entire output, potentially 400,000 iterations, is too large. We set the range to be from iteration 2,501 to 25,000, where we visually inspect that the former contains clusters too heterogeneous and the latter highly uniform clusters. In order to reduce noise, we plot the difference in conductance in Figure B1 between every hundred (rather than every two) iterations of the algorithm. We require that candidate solutions be about equally spaced so that our search will be relatively broad and not concentrate on a narrow area. Therefore, we split the search space for candidate solutions into six separate ranges ordered by number of iterations (2,501–5,000, 5,001–7,500, 7,501–10,000, 10,001–17,500, and 17,500–25,000) and then selected the highest peak in each range. We denote the five outcomes of this clustering routine as candidate solutions O_1, …, O_5 (labelled with capital O's to differentiate from the clustering outputs from the algorithm, which are labelled with small o's); they occur when (respectively) 3,129, 5,749, 9,690, 15,463, and 22,065 clusters have been identified.

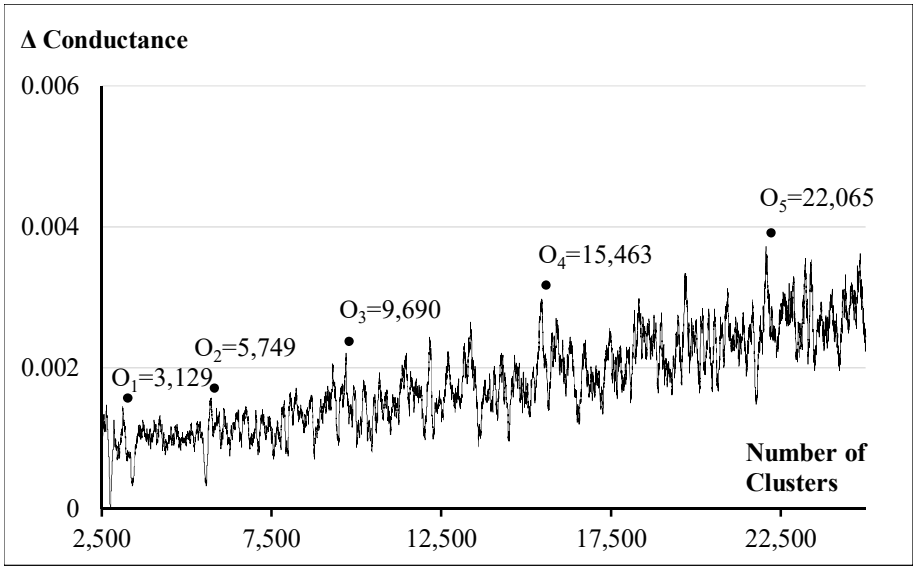


Figure B1: Tracking jumps in level of conductance between every 100 iterations

PART C: SURVEY FLOW

In all experiments subjects have to agree to the following consent form. Failing to agree to the conditions would automatically result in immediate termination of the survey. Boxes bounded are actual questions seen by subjects (each box represents a single screen).

CONSENT FORM

The purpose of this study is to improve our understanding of product designs.

To be eligible to complete the study and receive compensation, you must satisfy the following conditions:

- have resided in the US for at least 6 years
- be at least 18 years of age
- be a fluent speaker of English

If you satisfy the above conditions and would like to participate in this study, please click "Agree". Thank you for your participation.

- ○ Disagree
- ○ Agree

At the end of the survey, all subjects have to answer a verification question (given below).

This is a verification question. Please click on the button corresponding to the **chair**.

C.1 Replication of the Select Task

1. In order to answer the following questions, please note that objects belong to a style to the extent that they share a similar appearance.

In each of the following questions you will be shown 3 groups of product designs (10 designs in each group).

Your task is to identify the group that is the most heterogeneous with respect to styles. Note that a heterogeneous group would contain designs from different styles.

_____

Are the instructions clear?

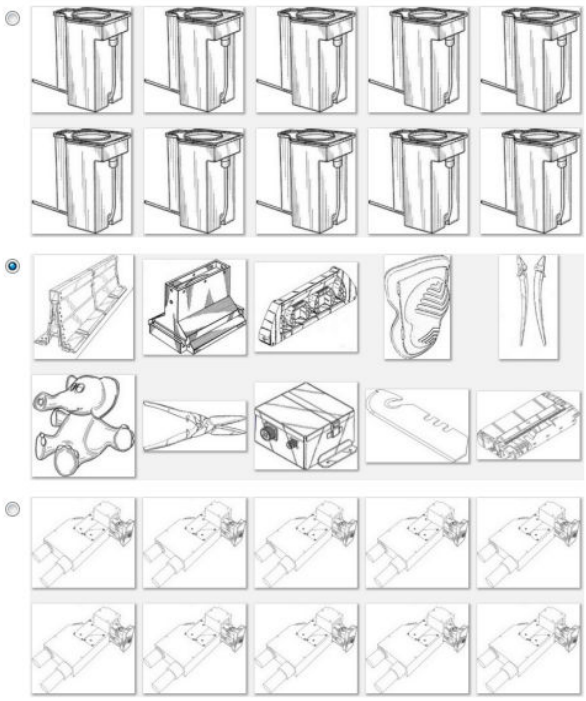
- Yes
- No

2. This section contains the questions to be analyzed. Please answer them openly and truthfully.

3. Identify the group that is the most heterogeneous with respect to styles. Note that a heterogeneous group would contain figures from different styles.

Item 3 are the actual tasks. A total of 10 questions are presented (randomly picked and presented in

4. Identify the group that is the most heterogeneous with respect to styles. Note that a heterogeneous group would contain figures from different styles.



random order).

Item 4 is a comprehension check question. A total of two questions of this type are presented.

C.2 Replication of the Partition Task

1. In order to answer the following questions, please note that objects belong to a style to the extent that they share a similar appearance.

In each of the following questions you will be shown a set of 10 product designs.

Your task is to categorize the designs into 2 groups (with each group containing exactly 5 designs), so that each group contains designs that share a similar appearance. That is, objects within each group look more similar to each other than with those in the other group.

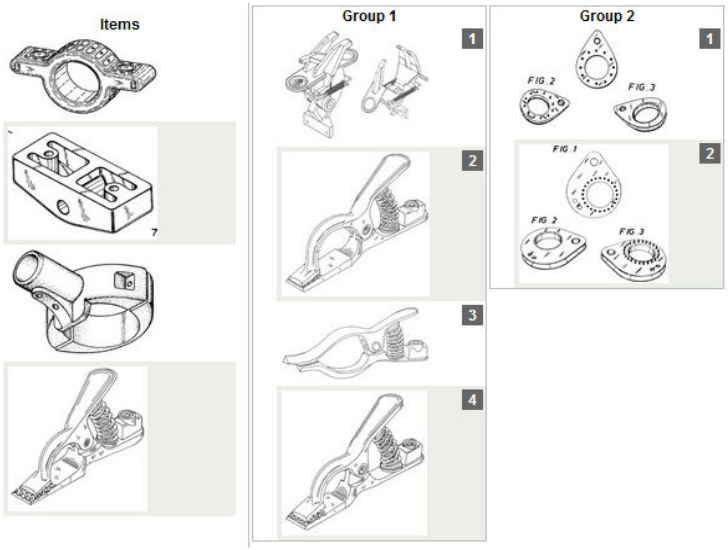
_____

Are the instructions clear?

- Yes
- No

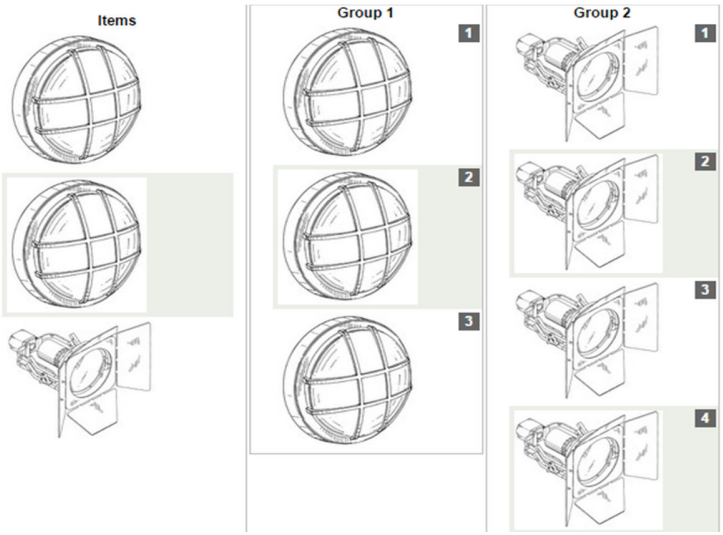
2. This section contains the questions to be analyzed. Please answer them openly and truthfully.

3. Categorize the designs into 2 groups (with each group containing exactly 5 designs), so that each group contains designs that share a similar appearance. That is, objects within each group look more similar to each other than with those in the other group.



Item 3 are the actual tasks. A total of 10 questions are presented (randomly picked and presented in random order).

4. Categorize the designs into 2 groups (with each group containing exactly 5 designs), so that each group contains designs that share a similar appearance. That is, objects within each group look more similar to each other than with those in the other group.



Item 4 is a comprehension check question. A total of two questions of this type are presented.

C.3 Turing Test (Select Task)

1. In order to answer the following questions, please note that objects belong to a style to the extent that they share a similar appearance.

In each of the following questions you will be shown 3 groups of product designs (10 designs in each group). The designs have previously been given to both a machine and a human, tasked to identify the group that is the most heterogeneous with respect to styles. Note that a heterogeneous group would contain figures from different styles.

In each question, we randomly select one of the outcomes of this selection exercise and present it to you. Your task is to decide whether this outcome is generated by a machine or a human.

_____

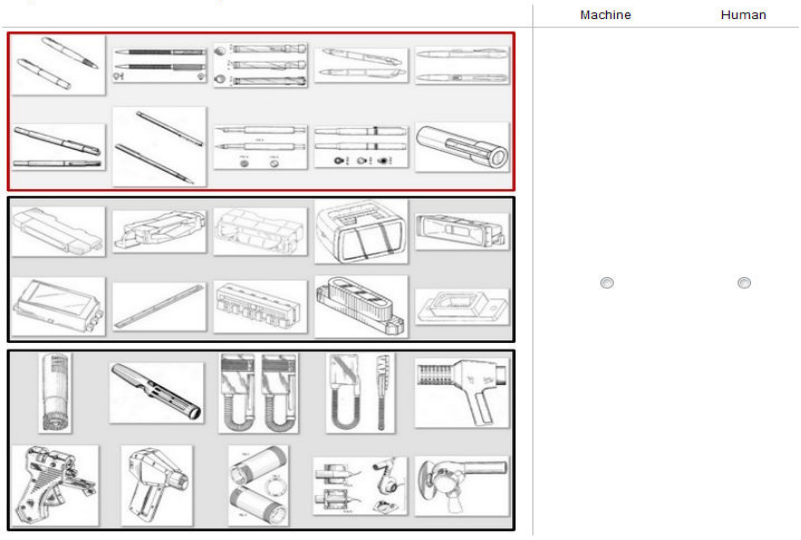
Are the instructions clear?

- Yes
- No

2. This section contains the questions to be analyzed. Please answer them openly and truthfully.

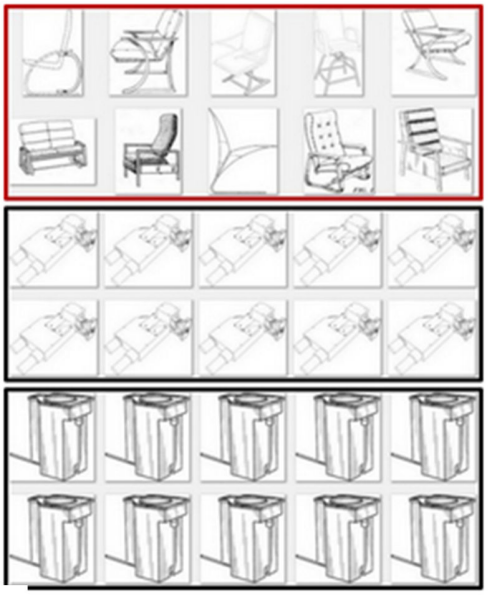
3. The group of designs at the top (in red borders) have previously been identified as the group that is the most heterogeneous with respect to styles. Note that a heterogeneous group would contain figures from different styles.

Do you think this is identified by a machine or a human?



Item 3 are the actual tasks. A total of 10 questions are presented (randomly picked and presented in random order).

4. Do you think the group of designs at the top (in red borders) is the most heterogeneous with respect to styles? Note that a heterogeneous group would contain figures from different styles.



- Yes
- No

Item 4 is a comprehension check question. A total of two questions of this type are presented.

C.4 Turing Test (Partition Task)

1. In order to answer the following questions, please note that objects belong to a style to the extent that they share a similar appearance.

In each of the following questions we will show a group of 10 product designs. The designs have previously been given to both a machine and a human, tasked to categorize the designs into 2 groups (with each group containing exactly 5 designs), so that each group contains designs that share a similar appearance. That is, objects within each group look more similar to each other than with those in the other group.

In each question, we randomly select one of the outcomes of this categorization exercise and present it to you. Your task is to decide whether this outcome is generated by a machine or a human.

_____

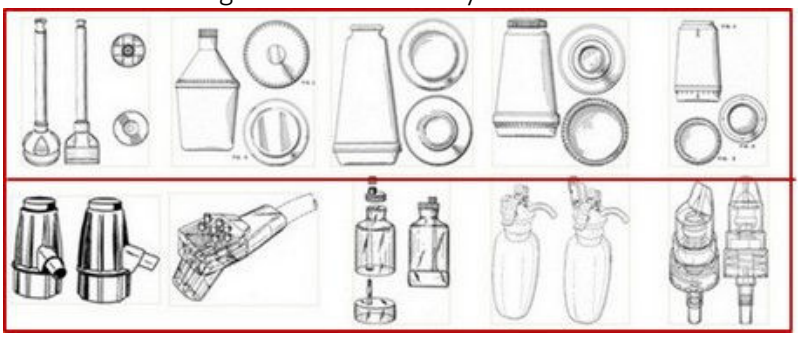
Are the instructions clear?

- Yes
- No

2. This section contains the questions to be analyzed. Please answer them openly and truthfully.

3. The following 10 designs have been categorized into two groups (the top row of 5 designs forming one group, and the bottom row of 5 designs the other group). The categorization is done so that each group contains designs that share a similar appearance. That is, objects within each group look more similar to each other than with those in the other group.

Do you think this categorization is done by a machine or a human?

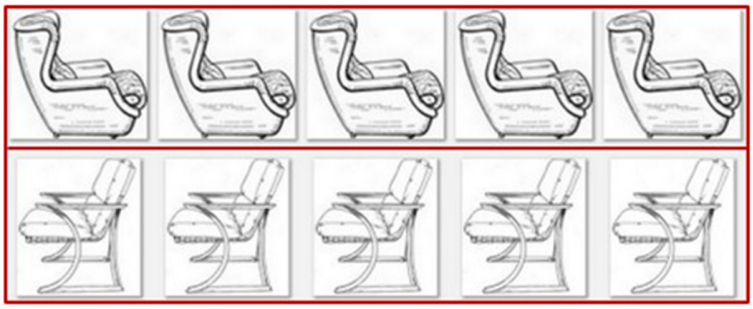


- Machine
- Human

Item 3 are the actual tasks. A total of 10 questions are presented (randomly picked and presented in random order).

4. The following 10 designs have been categorized into two groups (the top row of 5 designs forming one group, and the bottom row of 5 designs the other group).

Do you think the categorization is done so that each group contains designs that look, with respect to styles, more similar to each other than with those in the other group?



- Yes
- No

Item 4 is a comprehension check question. A total of two questions of this type are presented.

C.5 Evaluating the Candidate Solutions

1. In order to answer the following questions, please note that objects belong to a style to the extent that they share a similar appearance.

In each of the following questions we will show a group of 10 product designs. Please indicate whether you agree that the designs belong to the same style.

_____

Are the instructions clear?

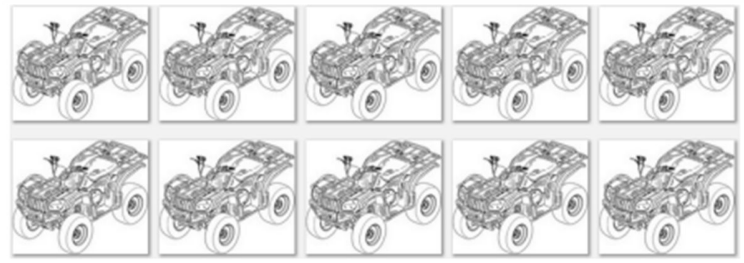
- ◯ Yes
- ◯ No

2. This section contains the questions to be analyzed. Please answer them openly and truthfully.

3. We show below a group of 10 product designs. Please indicate whether you agree that the designs belong to the same style.

Item 3 are the actual tasks. A total of 10 questions are presented (randomly picked and presented in random order).

4. We show below a group of 10 product designs. Please indicate whether you agree that the designs belong to the same style.



	Agree	Disagree
	○	○

Item 4 is a comprehension check question. A total of two questions of this type are presented.