THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Deciphering Clusters With a Deterministic Measure of Clustering Tendency

Citation for published version:
Diallo, AF & Patras, P 2024, 'Deciphering Clusters With a Deterministic Measure of Clustering Tendency', *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 4, pp. 1489-1501.
https://doi.org/10.1109/TKDE.2023.3306024

Digital Object Identifier (DOI):
10.1109/TKDE.2023.3306024

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE Transactions on Knowledge and Data Engineering

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

OPEN ACCESS

Download date: 11. May. 2024

Deciphering Clusters with a Deterministic Measure of Clustering Tendency

Alec F. Diallo and Paul Patras, *Senior Member, IEEE*

Abstract—Clustering, a key aspect of exploratory data analysis, plays a crucial role in various fields such as information retrieval. Yet, the sheer volume and variety of available clustering algorithms hinder their application to specific tasks, especially given their propensity to enforce partitions, even when no clear clusters exist, often leading to fruitless efforts and erroneous conclusions. This issue highlights the importance of accurately assessing clustering tendencies prior to clustering. However, existing methods either rely on subjective visual assessment, which hinders automation of downstream tasks, or on correlations between subsets of target datasets and random distributions, limiting their practical use. Therefore, we introduce the *Proximal Homogeneity Index (PHI)*, a novel and deterministic statistic that reliably assesses the clustering tendencies of datasets by analyzing their internal structures via knowledge graphs. Leveraging PHI and the boundaries between clusters, we establish the *Partitioning Sensitivity Index (PSI)*, a new statistic designed for cluster quality assessment and optimal clustering identification. Comparative studies using twelve synthetic and real-world datasets demonstrate PHI and PSI's superiority over existing metrics for clustering tendency assessment and cluster validation. Furthermore, we demonstrate the scalability of PHI to large and high-dimensional datasets, and PSI's broad effectiveness across diverse cluster analysis tasks.

Index Terms—Data Homogeneity, Clustering Tendency Assessment, Cluster Analysis, Knowledge Graphs, Knowledge Representation, Dimensionality Reduction, Exploratory Data Analysis

1 Introduction

Clustering algorithms are employed in a diverse range of machine learning (ML) and exploratory data analysis applications, including pattern recognition, computer networking, recommendation systems, market research, etc. [1]. Algorithms designed for this task seek to find intrinsic structures in data, allowing their separation into smaller groups of similar items based on some common characteristics. The utility of such information motivates its use as a pre-processing step for data analysis, or as dimensionality reduction methods to represent data with the most discriminative patterns [2], [3].

Despite the obvious utility of clustering algorithms, several fundamental aspects of clustering are still highly debated or overlooked. These include universally agreed upon definitions [4], choice of appropriate clustering algorithms that optimize the quality of partitions [5], and even elementary questions such as whether data is amenable to clustering [6]. Given the purpose of clustering algorithms, the first and most important question that should be answered is: *"Does a dataset contain any inherent grouping structure?"*

Existing clustering methods and their applications [7] are often impractical due to issues such as time and computational complexities, or sensitivity to outliers. Further, their imposition of a classification on a dataset, i.e., blindly forcing a partitioning of the dataset without prior knowledge of inherent structures, produces partitions regardless of whether any natural clusters are present, which can, and very often leads to misinterpretations. While the objective of a clustering task can be constructive (subject to intention) [8], the knowledge of inherent grouping structures still provides valuable insights allowing justification of choices or reconfiguration of datasets to highlight desired attributes. Hence, studying the *"clusterability"* or *"clustering tendency"* of datasets is a prerequisite for clustering.

Several clustering tendency assessment methods have been proposed over the years [9], [10], [11], [12]. These typically aim to determine the existence of meaningful clusters within a given dataset, thereby avoiding inappropriate clustering and misinterpretation of results. On one hand, discovering that a dataset does not possess sufficient cluster structure to be meaningfully partitioned indicates that clustering may not be suitable for the given data, or that the data may need to be reprocessed. Alternatively, if the data is found to be clusterable, a suitable algorithm can be selected or developed. Upon this assessment, the target dataset is clustered when fit and the quality of the resulting clusters is evaluated through clustering validation measures, which may trigger the selection of a clustering algorithm alternative until the produced clusters reach a desired quality. Skipping the clustering tendency assessment step, as observed in typical clustering pipelines, could potential lead to scenarios where considerable time and effort is wasted by applying countless clustering algorithms in an attempt to achieve good clustering results on data lacking cluster structures. This issue is overcome by initially ensuring that a dataset preprocessor produces representations suitable for clustering. The steps followed by these typical and ideal clustering pipelines are depicted in Fig. 1.

Deciding on a suitable assessment approach for clustering tendency is not straightforward. While prior methods for the evaluation of clusterability help in gaining insight into the behavior of clustering techniques, they all differ significantly and have practical limitations [6]. Specifically,

Alec F. Diallo and Paul Patras are with the School of Informatics, University of Edinburgh, Scotland, UK. (e-mail: {alec.frenn,paul.patras}@ed.ac.uk)

This work was partially supported by Arm Ltd.

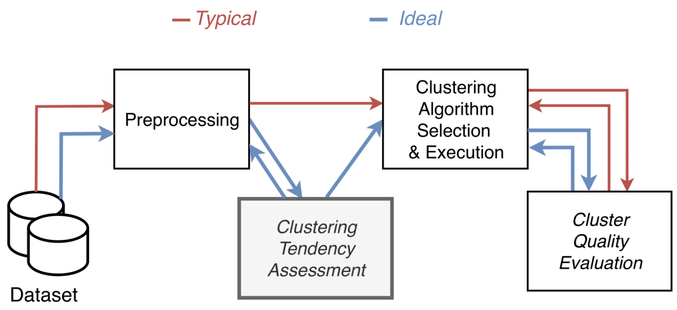

Typical
Ideal
Preprocessing
Clustering Algorithm Selection & Execution
Clustering Tendency Assessment
Cluster Quality Evaluation
Dataset

Fig. 1. Typical and ideal clustering pipelines, highlighting the possible succession of steps involved in each.

most fail to identify inherent structures in complex datasets or are not applicable in real settings due to their computational complexity. Further, many of the measures of clusterability are based on specific clustering algorithms or objective functions [13], [14], which effectively inverts the clustering pipeline (requiring that an algorithm is chosen before one determines whether data possesses sufficient structure to be meaningfully clustered), and thereby restricts the notion of clusterability to identifying structures that a chosen algorithm can capture.

In this paper, we propose the *Proximal Homogeneity Index* (PHI), a novel statistical scoring system that offers a deterministic and unbiased summary that describes the clustering tendency of a given dataset. Its formulation, designed to exploit global and local structures of inherent groups of data, produces a qualitative score that allows informed clusterability decisions, as well as the comparison of different topological configurations of samples. PHI, intuitively characterizing the homogeneity of samples, naturally provides a suitable tool for evaluating the quality of clusters discovered. Hence, we derive another statistic based on PHI, named the *Partitioning Sensitivity Index (PSI)*, which enables such evaluations solely relying on the relationships between samples of the dataset. Using PHI and PSI, we perform an extensive empirical analysis of data clusterability and cluster validations, and compare our results against those obtained with popular statistics to date, including the Purity [15] and Silhouette [16]) scores, demonstrating that our proposed statistics more often agrees with expert knowledge. We evaluate the effectiveness of our proposed statistics by addressing crucial aspects of exploratory data analysis, namely: *(i)* assessing the clustering tendency of a dataset; *(ii)* comparing different clustering or dimensionality reduction methods, to find the most appropriate one for a given dataset; and *(iii)* finding the optimal parameters (e.g., number of partitions) to be used by a clustering algorithm. Furthermore, we show that our approach, developed according to rigorous principles, scales well to the analysis of large, high-dimensional datasets.

The rest of the paper is organized as follows. We present our proposed statistics in Section 3 and perform extensive comparative evaluations in Section 4. Limitations and possible improvements of our approach are discussed in Section 5 before reviewing the relevant literature in Section 2, highlighting the shortcomings of existing methods and motivating the need for new cluster analysis tools such as the one proposed. Section 6 summarizes our findings and concludes the paper.

2 Related Work

Existing clusterability assessment methods can be categorized either as informal (graphical-based) or formal (statistics-based). With high-dimensional and complex datasets being ubiquitous in this era, very few of these methods have remained pertinent, and even fewer for multimodal applications.

Graphical-based tendency assessment methods [17] inspect the clusterability of datasets by generating visual forms which indicate the presence of different clusters in the set. They are often computed from randomly generated samples, which reduces their reliability and considerably degrades their performance when handling hierarchically ordered clusters. Due to their high memory and computational requirements [12], in addition to their inconclusiveness when faced with complex datasets, we only consider formal techniques in this study.

Viewing clustering tendency as a test for spatial randomness, existing statistics-based approaches are designed around the idea that random data should typically not have clusters [9], [18], [19]. However rare, this observation is not always true, which reduces the reliability of these methods. Intuitively, formal approaches traditionally measure the likelihood of a dataset's samples as being generated by a uniform sampling distribution, and therefore, use random sampling to compute their statistics, which ultimately makes their results vary across different evaluations, thereby requiring multiple evaluations to provide an average estimate. Their test for affinity towards aggregation relies on a null hypotheses such as H_0*: samples from the dataset are randomly distributed*. Another important issue with these methods is that since random sampling may fail to choose samples that provide a faithful representation of the dataset's intrinsic structures, their accuracy is subject to degradation, especially with large datasets. Hopkins and Skellam [9] proposed a statistic which compares the distances between randomly selected samples from a dataset and their nearest neighbors (w_i) within the dataset to distances between samples generated from a random distribution and their nearest neighbors within the dataset (u_i). This test, initially designed for 2-dimensional samples has been widely used after its extension to high-dimensional datapoints by Cross and Jain [18]: $H = \sum_i u_i^d / \Big(\sum_i u_i^d + \sum_i w_i^d \Big)$. With this definition, values close to 1 tend to suggest the presence clusters, while values near 0 tend to suggest uniformly distributed data. Randomly distributed data then tend to result in values around 0.5. While there is no definitive cut-off established for this statistic, values greater than 0.5 are usually considered as indicators for clusterable datasets. This statistic, which was found to be the most powerful [20], [21], has been observed to lose its effectiveness when used on high dimensional datasets [6]. Asides from the suboptimality of this approach due to its limitations on processing speed and size and dimensionality of datasets, its sensitivity to the number of samples randomly drawn and the fact that real datasets are never randomly uniform, make this statistic unfit for practical applications.

These shortcomings motivate our search for a novel, effective, and scalable clustering tendency assessment statistic, which unlike the Hopkins statistic, provides a deterministic score and a much simpler and more intuitive interpretation, with *a value of 1 representing a perfectly clustered dataset (homogeneous data that is not amenable to clustering) and a value of 0 indicating the presence of highly clusterable samples.* It is worth noting that since clustering tendency assessment statistics must be computed without a priori domain knowledge, it is extremely important for them to be consistent and reflective of the overall structure of datasets.

3 Proposed Statistics

The proposed Proximal Homogeneity Index (PHI) is a novel statistic that takes as input a set of data points and produces a score directly correlated to their structural homogeneity. This score, which is inversely proportional to the clustering tendency of the dataset, allows intuitive interpretations of its clusterability. That is, a target dataset is considered highly clusterable if its homogeneity index is low, and inversely, highly homogeneous datasets are considered unlikely to contain clusters.

PHI provides a deterministic score, as no assumptions are made about the topology of the dataset or the relationship between its samples and random distributions. This allows replicability of results across multiple runs, which makes our approach more reliable than existing methods relying on various null hypotheses based on randomization. Additionally, PHI is designed to be easily applicable to real-world datasets with no restrictions on size, dimensionality, or modality, and is completely independent of the clustering algorithms used to partition the data for downstream tasks.

Upon clustering, an accurate cluster validation metric (clustering quality index) is obtained from the PHI scores of individual clusters. This metric, which we call the *Partitioning Sensitivity Index (PSI)*, acts as a new statistic enabling several cluster validation tasks, such as finding the most appropriate clustering algorithm or the number of clusters that optimizes the overall homogeneity (cluster quality) of a partitioned set.

3.1 Clustering Tendency Assessment

Our proposed statistic computes the homogeneity index of a dataset by relying on measures of separability and compactness, which are given by a knowledge graph generated from the dataset's elements.

Suppose we have a collection of n vectors in a d-dimensional space, for which we wish to assess whether it contains any inherent grouping structures. To illustrate our approach, we will first consider samples in a 2-dimensional space, i.e., $\mathcal{S} := \left\{\mathcal{S}_i \in \mathbb{R}^2 \;\middle|\; i \in \{1, \cdots, n\}\right\}$.

Let us normalize the features of the set such that all samples lie within the space defined by a unit square. Using the Min-Max normalization, we ensure the preservation of relationships among the original samples, while the range of values is bounded as desired (between 0 and 1). This mapping of features from the original values contained in $\mathcal{S}$ to normalized values in $\mathcal{S}'$ is computed as follows:

$$\mathcal{S}' = \left\{ \frac{\mathcal{S}_i - \min \mathcal{S}}{\max S - \min \mathcal{S}} \;\middle|\; i \in \{1, \cdots, n\} \right\}. \tag{1}$$

To provide adequate localization properties while minimizing the computational complexity, we partition the space defined by a unit square into $n_p \times n_p$ disjoint subsets (grid partitions), where $n_p \in \mathbb{N}_{\geq 0}$ is the number of partitions along each dimension. Any sample is therefore identified to be in exactly one of the regions based on its coordinates (feature values); and all samples exactly at a decision boundary are assigned to their closest lower/left cell, with the exception of zero-valued features, which are assigned to the closest upper/right cell. Formally, let $C_{i,j} \in C$ define a cell of the partitioned grid:

$$\begin{aligned} C_{i,j} = &\left\{ \mathcal{S}'_k \;\middle|\; k \in \{1, \cdots, n\} \right\}, \text{ such that,} \\ &\frac{i}{n_p} < \mathcal{S}'^{(0)}_k \leq \frac{i+1}{n_p} \quad \text{or} \quad \mathcal{S}'^{(0)}_k = i = 0, \\ &\frac{j}{n_p} < \mathcal{S}'^{(1)}_k \leq \frac{j+1}{n_p} \quad \text{or} \quad \mathcal{S}'^{(1)}_k = j = 0, \end{aligned} \tag{2}$$

where $i, j \in \{0, \cdots, n_p - 1\}$, and $\mathcal{S}'^{(0)}_k$ and $\mathcal{S}'^{(1)}_k$ are the features of sample $\mathcal{S}'_k$ along the first and second dimensions, respectively.

A comprehensive summary (bird's-eye view) of the samples' distribution can then be obtained by replacing samples contained within each cell by their average. Let C' represent this summarized view of the dataset, with the content of each of its cells given by:

$$C'_{i,j} = \frac{1}{|C_{i,j}|} \sum_{k=0}^{|C_{i,j}|} (C_{i,j})_k, \tag{3}$$

where $|C_{i,j}|$ is the number of samples contained in cell $C_{i,j}$, and $(C_{i,j})_k$ is the k-th sample contained in the cell. These cell aggregations are performed for all non-empty cells, i.e., $C_{i,j} \neq \emptyset$. The knowledge graph of the dataset can be obtained using the adjacency matrix derived from this summarized view. Let X be the set of all non-empty cells of C', the adjacency matrix of the graph ($A \in \mathbb{R}^{N \times N}$) is then defined as:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots \\ \vdots & \ddots & \\ a_{N,1} & & a_{N,N} \end{pmatrix},$$

where $N = |X|$ being the number of non-empty cells in C'. Values of adjacency matrices are typically set such that $a_{i,j} \neq 0$ *iff* X_i and X_j are adjacent. To take the locality into account, we consider two cells to be adjacent if the distance between samples contained within them (i.e., $\|X_i - X_j\|$) is less than the diameter of a single cell. This constraint alone can however introduce biases towards densely packed regions, therefore, we also constrain each non-empty cell to only have connections to (be influenced by) its two closest samples. Since we partitioned our unit square space into $n_p \times n_p$ cells, each cell has a side-length of $1/n_p$ and a diameter of $(1/n_p)\sqrt{2}$. Based on these constraints, the adjacency matrix can then be computed as follows:

$$a_{i,j} = \begin{cases} \|X_i - X_j\| & \text{if } \; i \neq j, \; j \in 2_{NN}(X_i) \text{ and} \\ & \quad \|X_i - X_j\| < \frac{1}{n_p}\sqrt{2} \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $2_{NN}(X_i)$ represents the two nearest-neighbors of X_i (not including X_i itself). With this definition, the adjacency matrix preserves the locality of samples, while preventing biases towards dense local clusters. Yet, the possibility of having multiple graph components from such an adjacency matrix hinders the simplicity of studying the overall homogeneity structure of the dataset. To overcome this issue, we introduce the concept of Proximally-Connected Graphs (PC Graphs).

Definition (Proximally-Connected Graph)**.** *Let $\mathcal{G}(V, E)$ be a graph with a set of vertices V and a set of edges E, and let v_i and v_j denote any two distinct vertices in V. $\mathcal{G}$ is said to be proximally-connected if there exists an edge between vertices v_i and v_j only when at least one of the two following conditions is fulfilled:*

1) *v_i and v_j are considered adjacent;*
2) *removing $e_{i,j}$ disconnects $\mathcal{G}$ such that vertices of the two resulting components cannot be connected with edges of length (or weight) lower than that of $e_{i,j}$.*

Remarks. The definition of proximally-connected graphs introduces some useful and interesting properties, namely:

- Any proximally-connected graph is connected;
- Any connected sub-graph of a proximally-connected graph is proximally-connected;
- For a given set of samples, structural homogeneity is preserved across all possible proximally-connected graphs.

Using the graph $\mathcal{G}$ generated from our previously defined adjacency matrix A, we can create a PC graph by updating A such that all components of $\mathcal{G}$ are connected. Let V be the set of all vertices of $\mathcal{G}$. The adjacency matrix is then updated as:

$$a_{i,j} = \begin{cases} \|V_i - V_j\| & \text{if } \ V_i \in V^i, \ V_j \in \overline{V}^i, \text{ and} \\ & \|V_i - V_j\| = \min \|V^i - \overline{V}^i\| \ , \\ a_{i,j} & \text{otherwise} \end{cases} \qquad (5)$$

where $V^i = \cup \{V_k \mid \ k \in \{1, \cdots, i\}\}$ and $\overline{V}^i = V \setminus V^i$.

The graph generated from the updated adjacency matrix then satisfies all requirements of PC graphs, and therefore enables fast and structured exploration of the dataset. Fig. 2 depicts a summary of the steps involved in the generation of a PC graph from given samples.

With this representation of the data, we can observe that tightly grouped samples are characterized by short edges between vertices, and the separation of different groups is emphasized by longer edges. Conforming to PC graphs, the separation of different groups is measured based on their closest members, which is conceptually ideal for measuring the separation between arbitrarily shaped clusters.

Let us now denote the PC graph of the dataset as $\hat{\mathcal{G}}(V, \hat{E})$, where V and $\hat{E}$ are respectively, the sets of vertices and edges in $\hat{\mathcal{G}}$. Let $|e_{i,j}|$ be the length associated to $e_{i,j}$, the edge between vertices v_i and v_j. The Proximal Homogeneity Index of the samples in $\mathcal{S}$ can then be obtained by combining the compactness and separation information given by $\hat{\mathcal{G}}$:

$$\varphi_{\mathcal{S}} \ = \ \frac{1}{\max\limits_{e_{i,j}} |e_{i,j}|} \times \frac{1}{|\hat{V}|} \sum_{e_{i,j}} |e_{i,j}|, \qquad (6)$$

where $|\hat{V}|$ is the total number of vertices in graph $\hat{\mathcal{G}}$.

The score produced by this statistic directly represents the degree of structural homogeneity of the dataset. That is to say, when different grouping structures are present, their homogeneity is low (due to the difference between the maximum and average distances in the graph); and the absence of grouping structures is reflected by a marginal difference between the maximum and average distances, producing a high homogeneity score. Therefore, we consider a dataset clusterable if it has low homogeneity, i.e., there is a significant gap between inherent grouping structures.

The generalization of this approach to high-dimensional vector spaces can be tedious if naively applied. One such solution consists of computing PHI for every subspace formed by unique pairs of dimensions, and averaging the scores obtained to represent the homogeneity index of the entire dataset. While this might provide a more accurate descriptive value for the dataset, we opt for a more practical solution by directly computing the homogeneity index

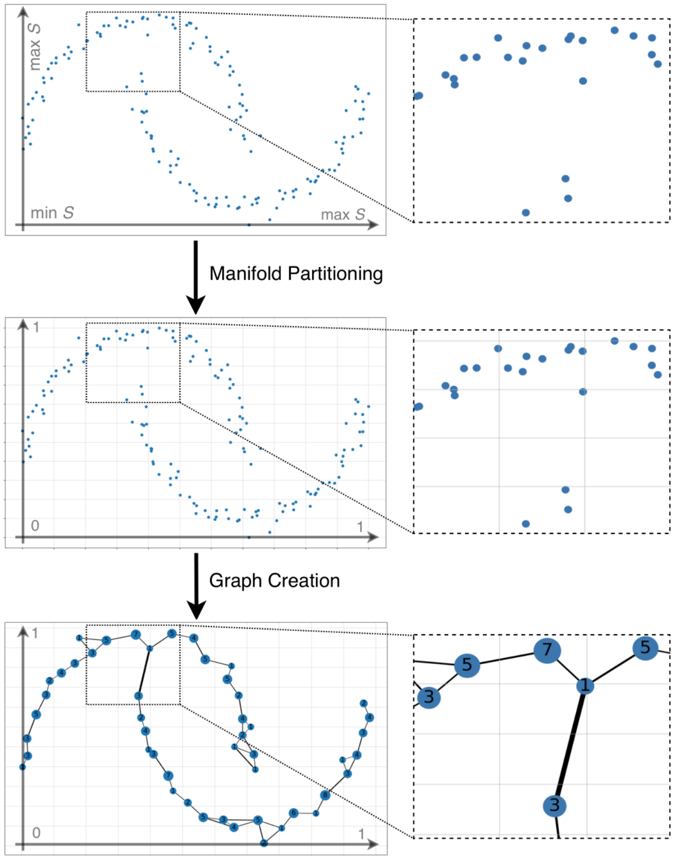


Fig. 2. Constructing a Proximally-Connected Graph from a dataset's samples. The top, middle, and bottom sub-figures show the original configuration of the samples, the re-scaled and grid-partitioned samples, and respectively the connected graph generated by the dataset. Each node's value corresponds to the number of samples it combines, and the distance separating two connected nodes is used as weight for the corresponding edge.

from the two most discriminative features of the dataset. This dimensionality reduction is achieved by using singular value decompositions [22] (SVD), where the new features are generated from weighted combinations of the original features. This not only improves the computational complexity of the model, but also has the advantage of being more robust to numerical errors, while providing additional privacy for sensitive, high-dimensional datasets through distortions and factorizations.

3.2 Cluster Validation

With the advent of big data, clustering is extensively used for exploratory data analytics. Accurately assessing the quality of the clusters obtained by an algorithm can be vital to achieving good performance in different applications. However, due to the unsupervised nature of the majority of tasks involving clustering, the performance of the downstream tasks cannot be evaluated based on a reserved evaluation set where the true clusters are known. Yet, most of the existing cluster validation metrics fail to accurately evaluate the quality of clusters without ground truth labels, which considerably hinders their practical viability.

Based on the proximal homogeneity index of the different clusters produced, we propose a new statistic to overcome these limitations. Specifically, this statistic, named the Partitioning Sensitivity Index (PSI), aims to accurately assess the quality of the clusters, without resorting to ground truth labels.

Given the clustering result obtained by an algorithm, the proximal homogeneity index φ_p is computed for every partition p found in the dataset. Let $\mathcal{P}$ denote the set of all partitions found by the clustering algorithm, we can summarize the clustering homogeneity (across all partitions) by comparing each partition to the least homogeneous partition in $\mathcal{P}$. However, to ensure this value is contained within the unit interval ([0, 1]), we invert the formulation such that the clusterability indices $(1 - \varphi_p)$ are used instead. The overall homogeneity across all partitions can thus be computed as follows:

$$\psi_P = \frac{1}{\max\limits_{p \in P}\,(1 - \varphi_p)} \times \frac{1}{|P|} \sum_{p \in P} (1 - \varphi_p), \qquad (7)$$

where $|P|$ is the total number of partitions produced by the clustering algorithm. For algorithms producing a single cluster (i.e., $|P| = 1$) or set of perfect clusters (i.e., $\max_{p \in P}\,(1 - \varphi_p) = 1$), the overall homogeneity is simply considered to be the average homogeneity score across all identified clusters.

Let $\mathcal{G}$ be the PC graph generated from P, and $\mathcal{G}_p$ the PC graph generated from partition p. And let ω_p denote the minimum distance separating vertices of p and vertices of $\{P \setminus p\}$. A partition p is considered correctly clustered if its longest edge is shorter than ω_p, i.e., all vertices of p are far away from any other partition. A vertex of p is considered correctly labelled if its distance to its closest neighbor in $\{P \setminus p\}$ is greater than the maximum length of its edges, i.e., the corresponding vertex is far enough from other partitions for p to be considered a clearly distinct partition.

Denoting as ρ_p the ratio of correct partitions, and ρ_v the ratio of correctly labelled vertices across all partitions produced by the clustering algorithm, the global correctness of the clustering can be derived as $\rho = \rho_p \cdot \rho_v$.

While we wish to penalize the overall clustering quality when the global correctness ratio is small, we adjust the severity of the penalty such that the clustering quality index deteriorates faster as the ratio gets smaller. This can be achieved by using a logarithmic function to set the trend of deterioration:

$$\overline{\rho} \;=\; \sqrt{\log_2\,(1 + \rho)} \qquad (8)$$

The normalization performed allows the penalty factor $\overline{\rho}$ to produce a value contained within the unit interval for any $\rho \in [0, 1]$. Using this adjusted penalty factor, the Partitioning Sensitivity Index (PSI) of the clustering is obtained:

$$\psi \;=\; \overline{\rho} \;\times\; \psi_P. \qquad (9)$$

This internal validation statistic, quantifying the quality of the clustering result based solely on the homogeneity of the clusters obtained, produces a value highly correlated to ground truth labels, and therefore effectively enables a wide range of downstream tasks as confirmed by the extensive evaluations we report in the next section.

4 Evaluation

We conduct several experiments to demonstrate the advantages of our proposed statistics over existing clustering tendency assessment and cluster validation metrics, while confirming that our statistics behave consistently across arbitrarily shaped datasets of different sizes and degrees of complexity. For all experiments conducted, we set the number of cells (grid partitions) n_p used to obtain the proximal homogeneity indices of datasets to $n_p = 2\,\lceil \ln(1 + n) - 1 \rceil$, where n is the number of samples of the dataset. Such value of n_p ensures computational efficiency on large datasets, by logarithmically (rather than linearly) increasing PHI's number of cells as the number of samples increases.

We design three sets of experiments focusing on: the clustering tendency assessment of datasets, the evaluation of clusters produced by an algorithm, and the scalability of our proposed approach. We use nine artificial and three real-world datasets for validation. We rely on the artificial datasets (depicted by the first two rows of Figure 3) to shed light on the strengths, weaknesses, and biases of the approaches evaluated, enabling clear and sound comparisons. We use real datasets primarily to highlight the practical viability of our proposed statistics, making the case for its adoption in statistical and exploratory data analysis.

4.1 Clustering Tendency Assessment

Our first experiment compares the scores produced by PHI to those produced by the Hopkins statistic. We refer interested readers to Section 2, where this baseline approach is discussed. Evaluations performed during this experiment with the artificial datasets cover a wide variety of scenarios, ranging from different number of clusters to intrinsic complexity of datasets (e.g., shapes of clusters, number and spatial disposition of samples). Each of these datasets was designed to represent specific characteristics used to compare the different statistics, namely proximity, randomness,

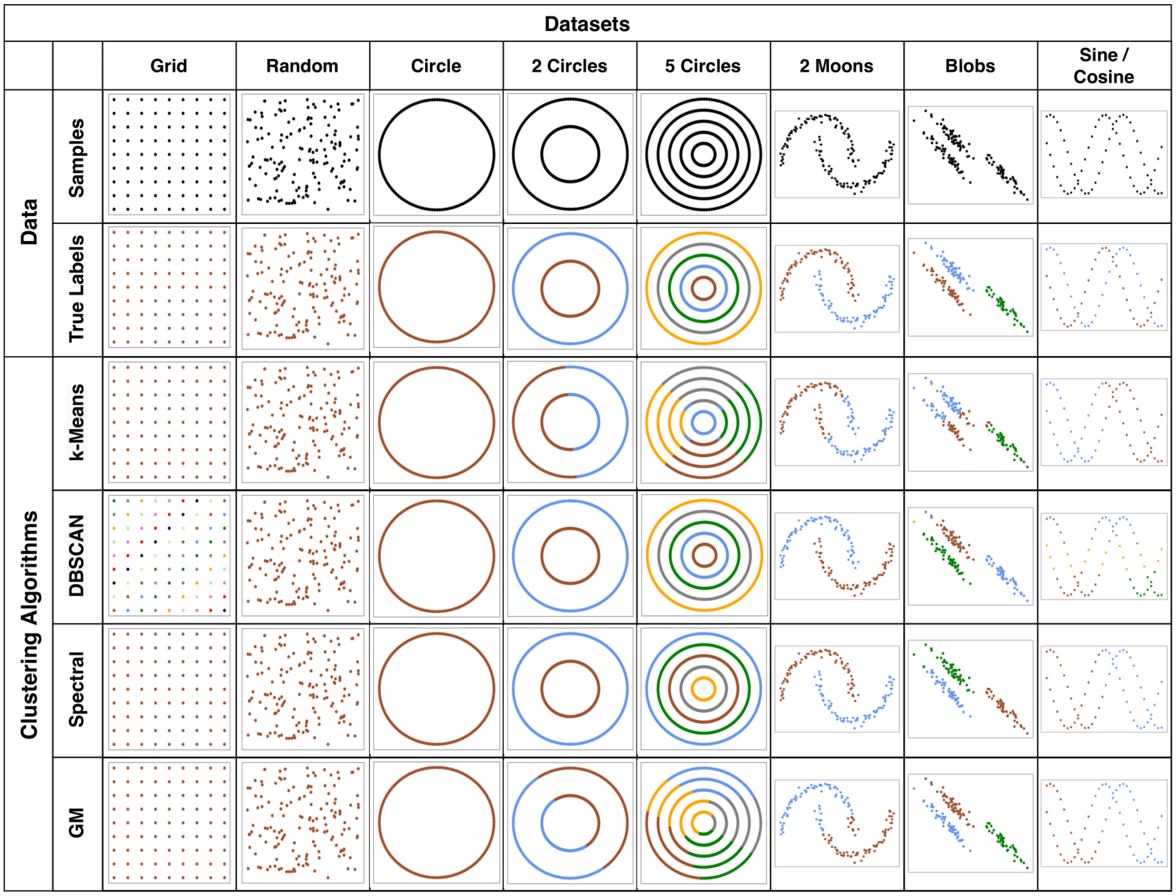


Fig. 3. Visual representations of our artificially generated datasets, and their clustering results by different algorithms (k-Means — *k-Means Clustering*, DBSCAN — *Density-Based Spatial Clustering of Applications with Noise*, Spectral — *Spectral Clustering*, and GM — *Gaussian Mixture*).

convexity, concentricity, affinity, interlacing, elongation, and overlap/intertwining. As per [19], [23], we use a sampling rate of 10% to calculate the Hopkins statistics. Although internal cluster validation metrics can indirectly measure clustering tendency, their need for pre-existing cluster assignments restricts their ability to evaluate the inherent clustering structure of raw, unclustered data. Consequently, these metrics fall outside the scope of the experimental results outlined in Table 1. Their separate evaluation will therefore be conducted in the following section (Section 4.2).

The results of our clustering tendency assessments are reported in Table 1, where clustering tendencies of samples are examined for each of the artificial datasets.

Revisiting the configuration of the datasets' samples (Fig. 3), we make the following observations about the clustering tendency assessments offered by the two statistics:

- The *Grid* dataset, considered perfectly homogeneous by PHI (i.e., not likely to contain any cluster), is according to the Hopkins statistic, randomly distributed.
- Both statistics tend to correctly indicate randomness when the samples are in fact randomly distributed.
- Where all circular shapes are considered by the Hopkins statistic likely to contain clusters (scores close to 1), PHI assesses their clusterability based on the homogeneity of samples and therefore produces low scores when distinct cluster structures are present.

TABLE 1
Evaluation of clustering tendencies using our proposed approach (PHI) and the Hopkins statistic. Averages and standard deviations are reported over 10 runs.

Dataset	Statistics	
	Hopkins	**PHI**
Grid	0.41 ±0.01	1.00 ±0.00
Random	0.64 ±0.05	0.59 ±0.00
1 Circle	0.87 ±0.02	0.83 ±0.00
2 Circles	0.82 ±0.02	0.30 ±0.00
5 Circles	0.75 ±0.01	0.67 ±0.00
2 Moons	0.80 ±0.03	0.36 ±0.00
Blobs	0.88 ±0.05	0.49 ±0.00
Sine / Cosine	0.58 ±0.05	0.75 ±0.00

- While the *5 Circles* dataset contains more clusters than the *2 Circles* dataset, PHI reports a higher homogeneity score for the former due to the *5 Circles* dataset having a more homogeneous overall structure than the the *2 Circles* dataset. This observation confirms PHI's ability to not only assess whether a dataset is clusterable, but also its intrinsic structure (i.e., how structurally homogeneous a dataset's samples are). For this comparison, the Hopkins statistic also reports a slightly lower clusterability.
- Structural patterns, completely ignored by the Hop-

kins statistic, constitute a key component of PHI, allowing more accurate assessments of clusterability and dataset homogeneity by PHI.

- Contrary to the Hopkins statistic that is based on randomness, PHI reports the same score across multiple runs due to its deterministic nature.

Overall, the results obtained demonstrate the viability of PHI as a clustering tendency assessment statistic, and further highlight its intuitive interpretation and consistency across different data representations.

4.2 Cluster Analysis

Having confirmed the effectiveness of PHI in assessing the clustering tendencies of datasets, we now wish to determine the suitability of its derived cluster validation statistic (PSI) for different cluster analysis tasks. Hence, based on the results of different clustering algorithms, we evaluate this statistic and compare its performance to those of common cluster validation metrics existing in the literature.

Since clustering validation metrics are typically either defined as *internal* (when they assess the quality of cluster structures without reference to external information), or *external* (when clusters are compared using prior or domain knowledge such as ground truth labels), we select 4 established approaches from each category as baselines to demonstrate the advantages of PSI, as follows.

External Cluster Validation Metrics

- **Purity score [15]:** measures the extent to which each cluster contains samples belonging to the most frequent label in that cluster. Its value ranges between 0 and 1, with larger values indicating better clustering.
- **Rand Index [24]:** measures the fraction of correctly clustered samples compared to ground truth labels. Its values (between 0 and 1) increase with clustering quality.
- **Adjusted Rand Index [25]:** corrects the Rand Index for chance, such that a baseline is established by using the expected similarity of clusters produced by a random model. While the Adjusted Rand Index can produce negative values (unlike the Rand Index), higher values still indicate better clustering.
- **Normalized Mutual Information [26]:** measures the amount of information shared between clusters produced by the clustering algorithm, and clusters defined by the ground-truth. Values range between 0 and 1, with larger values indicating better clustering.

Internal Cluster Validation Metrics

- **Silhouette Score [16]:** measures how well samples fit within their respective clusters. Its value ranges from -1 to 1, where 1 indicates well separated clusters, 0 indicates overlapping clusters, and negative values indicate samples assigned to the wrong cluster.
- **Dunn Index [27]:** compares the degrees of compactness and separation of clusters by dividing the minimum inter-cluster distance by the maximum cluster size. Larger values for this index indicate better clustering.
- **Davies–Bouldin Index [28]:** measures the average similarity of each cluster with its most similar cluster by calculating the ratio of within-cluster and between-cluster distances. With a minimum score of 0, lower values indicate better clustering.
- **Calinski-Harabasz Index [29]:** computes the ratio of total inter-cluster dispersions over all clusters and total intra-cluster dispersions over all clusters. Larger values for this index indicate better clustering.

4.2.1 *Comparing clustering algorithms*

In this experiment, we aim to find the most suitable clustering algorithm for any given dataset, based on the quality of clusters produced by candidate algorithms. To achieve this goal, we use our proposed statistic (PSI) to measure the performance of each candidate, and compare our selected choice against that made from ground truth labels. For each dataset used in this experiment, we use four clustering algorithms as candidates: k-Means [30], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [31], Spectral Clustering [32], and Gaussian Mixture models [33]. We implement these algorithms using Scikit-Learn library [34], keeping the default values of the relevant hyper-parameters, except for the number of cluster (where appropriate), which was specified according to each dataset. We use the same eight synthetic datasets and visualise the clusters determined by each of the algorithms in the lower four rows of Fig. 3. We evaluated the output of each algorithm by the 8 clustering validation baselines selected, our proposed statistic, and a true accuracy metric computed using the Hungarian algorithm [35], to find the best match between the clustering results and ground truth labels.

Table 2 compiles the results obtained for each dataset, clustering algorithm, and clustering evaluation metric. Based on the definitions of the clustering evaluation metrics (which values indicate better clustering), we compute the average number of times the best clustering algorithm was correctly indicated, i.e., the number of times a preferred algorithm had the maximum true accuracy across all candidates (bottom row – “Correct Algorithm Selection Ratio”). This evaluation allows an easy comparison of all the metrics considered. Since none of the internal clustering validation metrics used as benchmarks are applicable to datasets containing only one cluster, we consider them successful in comparing multiple clustering algorithms.

For each dataset, the results highlighted in the table show when the best performing clustering algorithm was indicated by the evaluation metrics. The visual representations of clustering results (depicted in Fig. 3), consulted in conjunction with Table 2, shows that: (i) the external clustering validation metrics used as baseline all ignore structural patterns (manifested by their failure to find the appropriate clustering algorithm for overlapping datasets), and (ii) none of the internal clustering validation metrics used as baseline is impervious to arbitrary configurations of clusters (e.g., elongation, concentricity, intertwinement).

While our proposed Partitioning Sensitivity Index (PSI) belongs to the category of internal clustering validation metrics (as we only rely on the results of clustering algorithms), we see in Table 2 that our correct algorithm selection ratio is higher than that made by all baselines, including

TABLE 2
Comparison of different clustering of datasets based on clustering validation metrics, and true (unbiased) accuracies computed using the Hungarian algorithm. Results are reported for our proposed statistic (PSI) and 8 existing validation metrics (Purity, RI — *Rand Index*, ARI — *Adjusted Rand Index*, NMI — *Normalized Mutual Information*, Sil. — *Silhouette score*, DI — *Dunn Index*, DB — *Davies-Bouldin Index*, and CH — *Calinski-Harabaz Index*).

Dataset	Clustering Algorithm	External Clustering Validation Metrics				Internal Clustering Validation Metrics					True Accuracy
		Purity	RI	ARI	NMI	Sil.	DI	DB	CH	PSI (ours)	
Grid	**k-Means**	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	DBSCAN	**1.00**	0.00	0.00	0.00	N/A	N/A	N/A	N/A	0.00	0.01
	Spectral Clustering	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	Gaussian Mixture	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
Random	**k-Means**	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	DBSCAN	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	Spectral Clustering	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	Gaussian Mixture	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
Circle	**k-Means**	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	DBSCAN	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	Spectral Clustering	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
	Gaussian Mixture	**1.00**	**1.00**	**1.00**	**1.00**	N/A	N/A	N/A	N/A	**1.00**	**1.00**
2 Circles	k-Means	0.50	0.50	0.00	0.00	0.35	0.01	1.18	173.16	0.64	0.50
	DBSCAN	**1.00**	**1.00**	**1.00**	**1.00**	0.11	**0.25**	449.98	0.00	**1.00**	**1.00**
	Spectral Clustering	**1.00**	**1.00**	**1.00**	**1.00**	0.11	**0.25**	449.98	0.00	**1.00**	**1.00**
	Gaussian Mixture	0.50	0.50	0.00	0.00	0.35	0.01	1.18	172.17	0.86	0.50
5 Circles	k-Means	0.40	0.75	0.23	0.28	0.32	0.01	0.90	490.76	0.57	0.22
	DBSCAN	**1.00**	**1.00**	**1.00**	**1.00**	-0.08	**0.10**	989.97	0.00	**0.67**	**1.00**
	Spectral Clustering	**1.00**	**1.00**	**1.00**	**1.00**	-0.08	**0.10**	989.97	0.00	**0.67**	**1.00**
	Gaussian Mixture	0.34	0.69	0.07	0.13	0.30	0.01	0.98	434.67	0.43	0.23
2 Moons	k-Means	0.75	0.63	0.25	0.19	0.49	0.05	0.77	224.21	0.85	0.73
	DBSCAN	**1.00**	**1.00**	**1.00**	**1.00**	0.32	**0.14**	1.17	95.20	**0.95**	**1.00**
	Spectral Clustering	**1.00**	**1.00**	**1.00**	**1.00**	0.32	**0.14**	1.17	95.20	**0.95**	**1.00**
	Gaussian Mixture	0.83	0.72	0.44	0.35	0.47	0.01	0.80	199.06	0.64	0.81
Blobs	k-Means	0.78	0.79	0.52	0.56	0.51	0.05	0.71	364.36	0.53	0.60
	DBSCAN	1.00	0.98	0.96	0.95	0.44	0.13	0.68	155.05	0.48	0.94
	Spectral Clustering	0.98	0.97	0.94	0.92	0.46	0.01	0.88	264.95	0.49	0.96
	Gaussian Mixture	**1.00**	**1.00**	**1.00**	**1.00**	0.45	0.11	0.96	251.53	**0.59**	**1.00**
Sine / Cosine	**k-Means**	0.50	0.49	-0.02	0.00	**0.57**	0.06	**0.58**	**162.15**	**0.73**	**0.50**
	DBSCAN	0.60	0.51	0.00	0.07	0.24	0.04	1.85	30.18	0.25	0.26
	Spectral Clustering	0.51	0.49	-0.01	0.00	0.57	0.06	0.58	161.13	0.71	0.50
	Gaussian Mixture	0.54	0.50	-0.01	0.01	0.56	0.07	0.59	144.68	0.71	0.47
Correct Algorithm Selection Ratio		0.88	0.88	0.88	0.88	0.50	0.75	0.50	0.50	**1.00**	

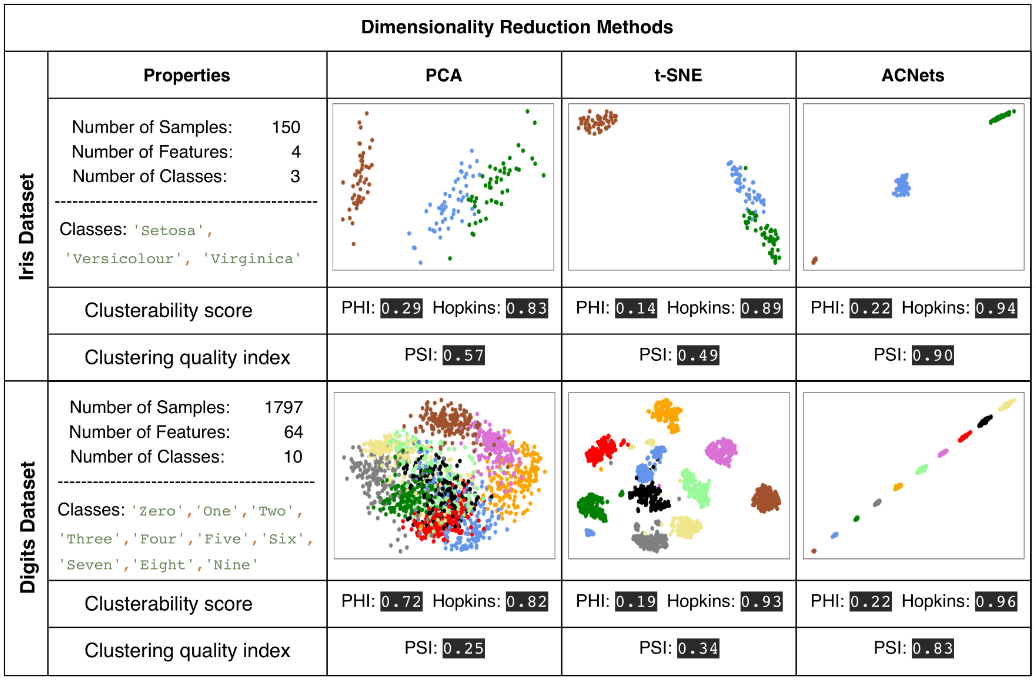
those of external metrics. By successfully finding the most appropriate algorithm for all evaluated datasets, without recourse to ground-truth labels, PSI shows great promise for significantly reducing the need for labelled datasets.

4.2.2 Comparing dataset representations

Given the importance of visual representations of data (providing a clear idea of how the information is partitioned and how different partitions relate to each other), we design an experiment focused on finding the right spatial configuration of a dataset's samples, so as to maximize the separation of different clusters. For this experiment, we used two realistic datasets widely popular in the pattern recognition domain, namely the Iris plants [36] and the hand-written Digits [37] datasets (see Fig. 4).

We use three different dimensionality reduction methods to generate candidate representations of the datasets in a two-dimensional space: Principal Component Analysis (PCA) [38], t-distributed Stochastic Neighbor Embedding (t-SNE) [39], and Adaptive Clustering networks (ACNets) [3]. On one hand, PCA, one of the most widely used dimensionality reduction methods, was designed to generate embeddings that retain most information about the dataset. On the other hand, t-SNE was designed to preserve local similarities while generating low-dimensional samples suited for visualization. By including ACNets in our benchmark, we considered a third scenario consisting of optimal separation of different clusters based on domain knowledge.

Using our proposed Partitioning Sensitivity Index, we can then compare the candidates obtained to find representations of the datasets that best separate different clusters. Fig. 4 shows for each dataset, the embeddings obtained by each method as well as their associated qualities according to PSI. From this figure, we observe that embeddings generated by ACNets have better partitioning qualities than those of PCA and t-SNE. Additionally, we see that while the embeddings generated by t-SNE are in general better separated than those produced by PCA, individual clusters generated



Dimensionality Reduction Methods				
	Properties	PCA	t-SNE	ACNets
Iris Dataset	Number of Samples: 150 Number of Features: 4 Number of Classes: 3 Classes: 'Setosa', 'Versicolour', 'Virginica'			
	Clusterability score	PHI: 0.29 Hopkins: 0.83	PHI: 0.14 Hopkins: 0.89	PHI: 0.22 Hopkins: 0.94
	Clustering quality index	PSI: 0.57	PSI: 0.49	PSI: 0.90
Digits Dataset	Number of Samples: 1797 Number of Features: 64 Number of Classes: 10 Classes: 'Zero','One','Two', 'Three','Four','Five','Six', 'Seven','Eight','Nine'			
	Clusterability score	PHI: 0.72 Hopkins: 0.82	PHI: 0.19 Hopkins: 0.93	PHI: 0.22 Hopkins: 0.96
	Clustering quality index	PSI: 0.25	PSI: 0.34	PSI: 0.83

Fig. 4. Representations obtained by different dimensionality reduction methods for the Iris and Hand-Written Digits datasets. Below each representation are the clustering tendency assessment (clusterability) scores obtained by PHI and the Hopkins statistic, and the clustering quality scores indicating how well separated the clusters are according to our proposed Partitioning Sensitivity Index (PSI).

by t-SNE are also more compact, which often creates better partitioning (seen with the hand-written Digits dataset), but can also combine different clusters when they are difficult to separate (as seen with the Iris dataset). Nevertheless, we see that in all cases, PSI accurately describes clustering qualities.

Upon further analysis, we highlight different clustering tendencies for the different clusters obtained. PHI and the Hopkins statistic are used to evaluate the clusterability of the embeddings generated by PCA, t-SNE, and ACNets for the two datasets. Where the Hopkins statistic indicates high clustering tendency for all of these representations (values close to 1), we see that our proposed statistic describes their clustering tendencies with high fidelity, i.e., low values where multiple clusters are present and high values where distinct clusters are hard to dissociate from others.

4.2.3 *Optimizing clustering algorithms*

Aside from knowing which clustering algorithm would best partition a dataset, another important question often encountered in cluster analysis is how to find the best hyper-parameters that maximize the performance of a given clustering algorithm. To answer that question, we compare the quality of partitions produced for each candidate value of the hyper-parameter considered. We use our proposed cluster validation statistic PSI as evaluation metric, and as target dataset the ACNets embeddings of the Iris samples. Since determining the optimal number of clusters constitutes a fundamental issue in clustering problems where the number of clusters is required to be set manually, we choose this as the hyper-parameter to optimize. As such, PSI is evaluated on three different clustering algorithms that require this. For completeness, we include two other clustering validation statistics to compare PSI against: Silhouette score and Calinski-Harabaz index.

Fig. 5 plots the cluster validation scores obtained for each number of clusters and each clustering algorithm considered. We additionally used the Consensus Clustering [40] algorithm to report the optimal number of clusters found across multiple runs of each algorithm, which is shown as the vertical dotted line on the plot, indicating 3 as the optimal number of clusters. For all clustering algorithms used, the scores obtained by PSI are validated by the Silhouette score, the consensus clustering algorithm, and the domain knowledge (i.e., known number of clusters — see the properties of the Iris dataset in Fig. 4). The Calinski-Harabaz index, however, fails to find the optimal number of clusters, or even decisively choose a best performing algorithm. This is explained by the formulation of this index, i.e., with clusters produced by ACNets, samples are tightly packed together and therefore dividing inter-cluster distances by intra-cluster distances often yields very large values, even when some clusters are not well separated. The results of this evaluation further highlight some key differences between our statistics and existing methods, mainly due to the ability of our approach to consider the structure of cluster samples in its definition. In this experiment, we observe that while the Silhouette scores allow for identification of the optimal number of clusters to be used, their values do not entirely reflect the quality of the clustering results. Namely, the Silhouette Score assumes that clusters are convex and isotropic, which significantly reduces its descriptive power when used to analyze complex or random cluster structures. As an internal cluster validation statistic, PSI is not only applicable under the same conditions as existing internal validation metrics, but often solves shortcomings of these methods as shown by Fig. 5.

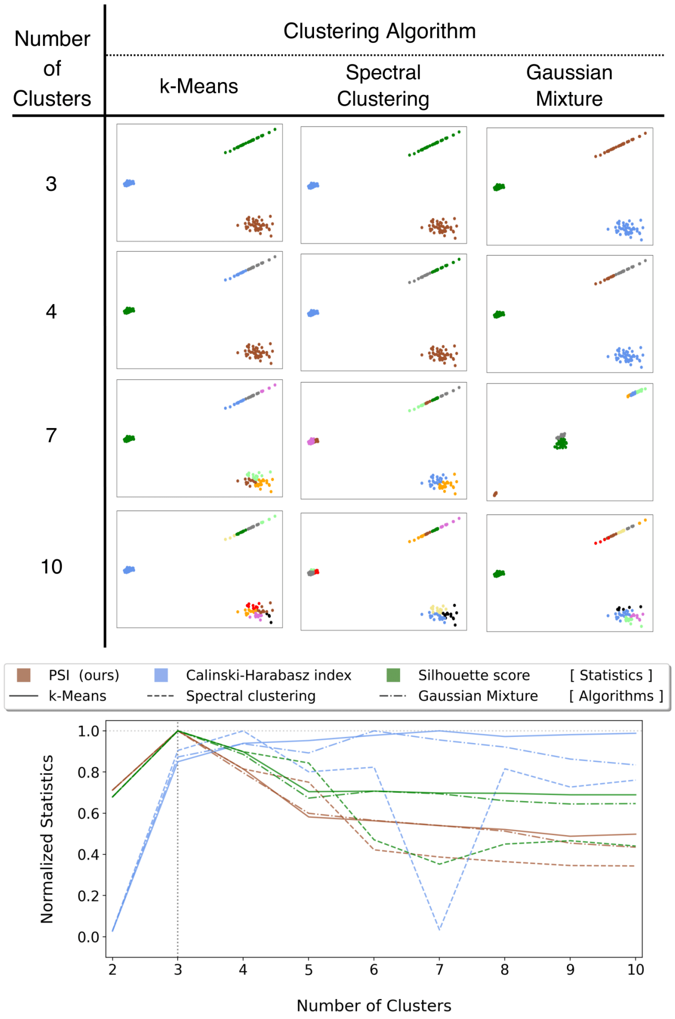


Fig. 5. Comparison of cluster qualities for clusters obtained by different partitioning algorithms, for multiple values of a single parameter (number of clusters). The top sub-figure shows visual representations of the clustering results (sampled for different number of clusters), and the one at the bottom plots the values recorded for all parameters. The number of clusters obtained by the consensus clustering algorithm is plotted as the vertical line.

4.3 Practical Use-Case

Next, we evaluate the performance of our proposed Partitioning Sensitivity Index when applied to large, complex, and challenging datasets. For this experiment, we use three realistic datasets, often used in the cyber-security and image recognition domains, namely CIC-IDS-2017 [41], MNIST [42] (a well-known handwritten digit classification dataset), and CIFAR-10 [43] (an established computer-vision dataset used for object recognition).

Results shown in Fig. 6 follow the changes in data representations and in PSI, as ACNets learns to separate the different clusters in the datasets. Depending on the difficulty of the learning task, we see that representations take between 60 and 200 training iterations to optimize. The quickest optimization achieved is for the CIC-IDS-2017 dataset, which contains 34,220 samples of 80 network traffic features, representing the most common cyber attacks known today. Comparatively, MNIST and CIFAR-10 were the second and third to have optimized representations. These datasets consist of 70,000 samples of 784 features and 60,000 samples of 3072 features, respectively.

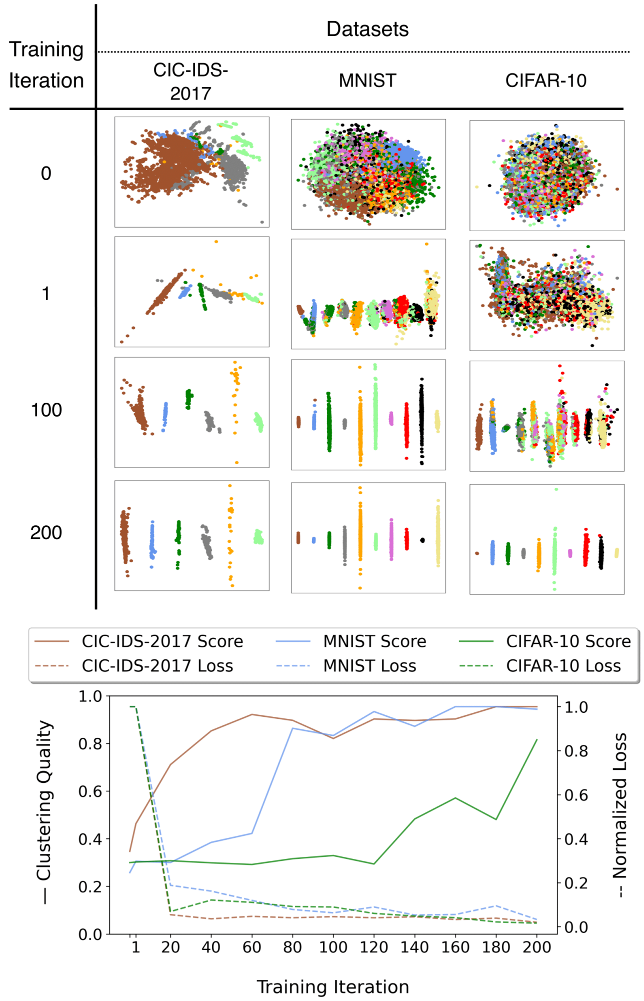


Fig. 6. Measures of cluster qualities at different stages of the training process. The size and complexity of the datasets are reflected by PSI's rate of improvement, as confirmed by the the qualities of embeddings produced.

An interesting observation from this experiment is that compared to the training loss, PSI better represents the actual performance of the neural networks. This is indicated by having the loss values close to their optimal while the representations are still far from optimized (e.g., CIFAR-10 representation, loss, and clustering quality at iteration 100).

4.4 Complex Datasets

In this experiment, we evaluate the robustness of our proposed cluster validation statistic on complex datasets with large number of clusters. The Aggregation dataset [44] consisting of seven perceptually distinct groups of points is used, as this contains features that are known to create difficulties for clustering algorithms such as: narrow bridges between clusters, uneven-sized clusters, and so on. Our proposed cluster validation statistic PSI is then used to evaluate the quality of clusters in this dataset. To showcase its robustness and faithful descriptive performance of the cluster qualities, we apply our approach to the dataset with all features that render clustering difficult, and to two variations of the dataset where some samples are selectively removed to simplify clustering.

The results presented in Fig. 7 highlight the ability of PSI to handle multiple clusters of complex shapes and sizes, and show that our approach accurately describes the quality of clusters obtained. In the left column of the figure, the original dataset, with narrow or interconnected clusters is evaluated by PSI to produce a quality index of *0.5*, where removing samples connecting different clusters significantly increase the quality index. Specifically, PSI obtains a score of *0.74* when only two clusters are bridged by samples, and a score of *0.82* when all clusters are well separated. This not only confirms the robustness of PSI in evaluating large number of clusters, but shows how the structural relationships between samples (see bottom row of the figure) impact the quality of clusters.

4.5 Scalability Analysis

While the majority of modern computing infrastructures are designed for scalability, throughput, and resilience, they often delegate the task of latency management to algorithmic tools. However, with the exponential growth of data availability, datasets are continuously increasing in size, dimensionality, and complexity, forcing data intensive applications to introduce a trade-off between scalability and functional accuracy [45], [46].

Exploratory data analysis techniques, having to analyze, understand, describe, and eventually extract useful information from their inputs, are most susceptible to requiring such a trade-off when presented with such large datasets. Therefore, we study the scalability of our proposed clustering tendency assessment statistic (PHI) and compare it to that of the Hopkins statistic (the baseline method discussed earlier). For this experiment, we measure the time taken to compute with each statistic, the clustering tendency of a randomly generated dataset. The size of the datasets are gradually increased to also report the growth rates for each method. Fig. 8 shows the average execution times over 10 runs, for each of the statistics considered.

Like many other cluster analysis techniques, the Hopkins statistic suffers from severe scalability problems as it is greedy in nature. Further, it is subjected to a trade-off between scalability and accuracy, as sampled data used to compute its clusterability scores are randomly selected, i.e., despite not being scalable, its accuracy still suffers due to randomness. Therefore, selecting all samples to compute a deterministic value for the Hopkins statistic would further degrade the scalability of the statistic.

Despite only using a 10% sampling rate to compute the Hopkins statistic (as recommended in the literature), a significant difference can be immediately noticed between the scalability of our PHI statistic and that of the Hopkins statistic. A dataset containing $10,000$ samples of $1,000$ features each has its clustering tendency measured by the Hopkins statistic in 60 seconds (using only 10% of the data), while PHI takes less than 1 second (using all samples), leading to a 59× improvement in processing speed. With respect to the dataset size, where the execution time increases exponentially for the Hopkins statistic, the growth rate of PHI is sublogarithmic. This can theoretically be proven by analyzing the complexity of all components involved in the computation of PHI. Assuming the worse case scenario, where all samples of the dataset are evenly distributed across the all partitions, and with n_p being the total number of partitions, each cell can have at most 8 neighbours, i.e. 9 cells considered for each sample. Using a PC graph to compute PHI, pairwise distances between aggregated cell samples can be computed in $\mathcal{O}(n_p^2 \ln n_p^2)$ using performance-oriented algorithms such as kd-Tree [47]. This term constitutes the most expensive operation as the construction of the graph itself is directly proportional to the number of edges, i.e., at worst $\mathcal{O}(n_p^2)$. As the value of n_p proposed in our experiment is logarithmic in terms of number of samples in the dataset ($n_p = 2\lceil \ln(1+n) - 1 \rceil$), the overall complexity of PHI can be reduced to $\mathcal{O}(\text{PHI}) = \mathcal{O}((\ln^2 n) \times (\ln(\ln^2 n))) = \mathcal{O}(k^2 \ln k)$, where $k = \ln(n)$. However, for sub-optimal pairwise distance calculation methods, the worse complexity of PHI, assuming the same value for n_p would be $\mathcal{O}(\text{PHI}) = \mathcal{O}(\ln n)$. This highly efficient complexity of PHI is obtained due to the main attributes of our approach (as discussed in Section 3.1), namely:

- the discriminative features' extraction via SVD,
- the input space discretization via grid partitioning,
- the logarithmic growth rate of number of grid cells,
- the combination of samples within each grid,
- and the Proximally-Connected Graph of the dataset.

4.6 Code Availability

Upon publication, the source code of our proposed approach will be made public on GitHub.

5 Discussion

Our proposed statistics, used as described in Section 3 perform distinctively well in all evaluation scenarios considered. We note however the existence of some edge cases, such as evaluating on sets containing only one or two samples, or understanding how outliers should participate to the final scores. Similarly, the value of n_p, indicating how many cells PHI should use for a given dataset, was set via trial and error to estimate the best number of grid partitions based on the number of samples. This parameter mainly enables faster processing of large datasets by aggregating local groups of samples into single data points while preserving the overall structure of the dataset. This can however introduce a trade-off between the speed of execution and the fidelity of the generated knowledge graph. As such, large values of n_p are to be avoided to prevent inaccurate assessments of the clustering tendency of datasets. Ultimately, while our formulation of n_p (Section 4) works consistently well across all evaluations performed, more theoretically-grounded definitions should be studied in future work.

Although our proposed statistics have overcome limitations of existing approaches, the structural patterns described by PHI do not currently take into account the number of samples merged by each cell. This can however be easily addressed (if needed) by adjusting PHI to account for

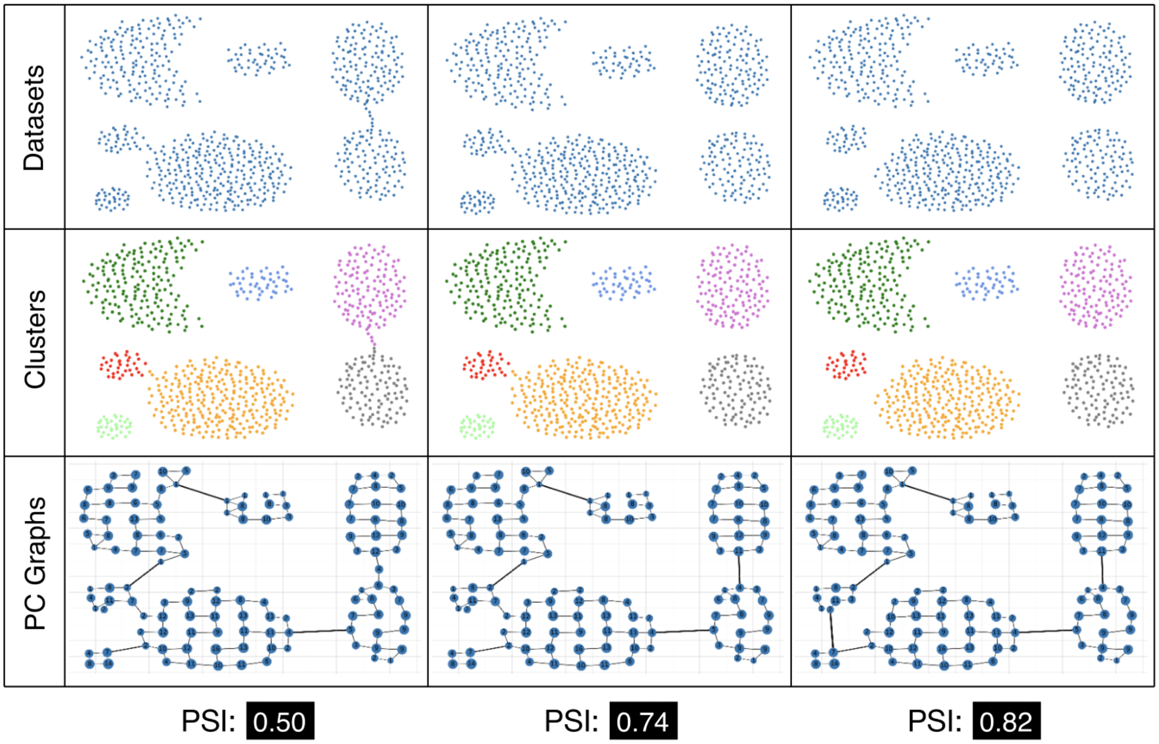


Fig. 7. Assessment of performance robustness of PSI on intricate cluster configurations. The clustering quality of the Aggregation dataset is evaluated under three distinct conditions – the original dataset (left column), and two modified versions with select and progressive sample removals to enhance cluster separation (center and right columns). Top and middle rows depicts the spatial configuration of the samples and are color coded to indicate their corresponding clusters, respectively. The bottom row shows the *PC graphs* generated from the samples and used to compute PSI.

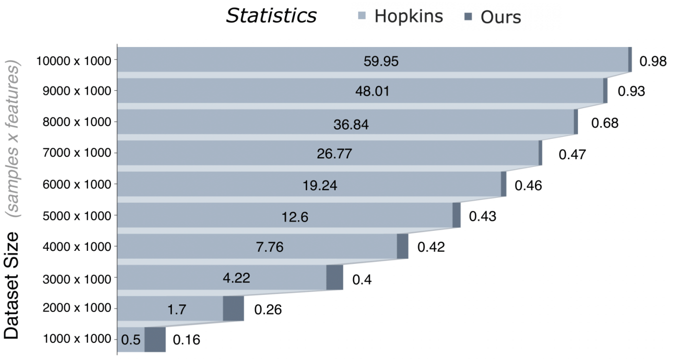


Fig. 8. Scalability analysis of PHI. Randomly generated datasets are used to compare the execution time of PHI to that of the Hopkins statistic (plotted on logarithmic scale and displayed in seconds), reporting averages over 10 runs. For each dataset, the Hopkins statistic evaluated using 10% of the samples.

disparities between the numbers of cells combined across all connected vertices.

As shown in the previous section, our approach painlessly scales to very large datasets. Therefore, we defer to future work further reductions of the execution time by using parallel processing techniques for 1) computing portions of the adjacency matrix simultaneously, and 2) enabling local exploration of extremely large graphs.

Finally, we leave for future work the application of our proposed statistics to unsupervised machine learning tasks.

6 Conclusion

Although clustering tendency assessment is generally overlooked by the research community, its undeniable necessity has become more apparent with the emergence of big data and machine learning applications. In this paper we have shown how such methods can be beneficial for exploratory data analysis applications and argued for the need of consistent, reliable, and deterministic statistics such as the Proximal Homogeneity Index (PHI) proposed. Through extensive experiments, we have demonstrated the suitability of PHI not only for clustering tendency assessment, but also, through our Partitioning Sensitivity Index (PSI), its effectiveness in validating clustering results, finding the most appropriate clustering algorithm and hyperparameters, and choosing the best dimensionality reduction method to maximize clustering performance. We studied PHI's practical viability using realistic datasets and revealed its scalability properties. The results obtained confirmed the consistency of our proposed statistics across datasets and clustering methods, and we made the case for their adoption in statistical and exploratory data analysis.

References

[1] Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on neural networks*, 16(3):645–678, 2005.

[2] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm sigkdd explorations newsletter*, 6(1):90–105, 2004.

[3] Alec F Diallo and Paul Patras. Adaptive clustering-based malicious traffic classification at the network edge. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10, Anchorage, AK, USA, 2021. IEEE, IEEE.

[4] Vladimir Estivill-Castro and Jianhua Yang. Fast and robust general purpose clustering algorithms. In *Proceedings of the 6th Pacific Rim International Conference on Artificial Intelligence*, PRICAI'00, page 208–218, Berlin, Heidelberg, 2000. Springer-Verlag.

[5] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.

[6] Andreas Adolfsson, Margareta Ackerman, and Naomi C Brownstein. To cluster, or not to cluster: An analysis of clusterability methods. *Pattern Recognition*, 88:13–26, 2019.

[7] Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. *Introduction to Data Mining (2nd Edition)*. Pearson, USA, 2nd edition, 2018.

[8] Christian Hennig. What are the true clusters? *Pattern Recognition Letters*, 64:53–62, 2015.

[9] Brian Hopkins and John Gordon Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(2):213–227, 1954.

[10] Trevor F Cox and Toby Lewis. A conditioned distance ratio method for analyzing spatial patterns. *Biometrika*, 63(3):483–491, 1976.

[11] James C. Bezdek and Richard J. Hathaway. Vat: A tool for visual assessment of (cluster) tendency. *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, 3:2225–2230, 2002.

[12] Jacalyn M Huband, James C Bezdek, and Richard J Hathaway. bigvat: Visual assessment of cluster tendency for large data sets. *Pattern Recognition*, 38(11):1875–1886, 2005.

[13] Xianchao Zhang and Quanzeng You. Clusterability analysis and incremental sampling for nyström extension based spectral clustering. *IEEE Intl Conference on Data Mining*, 11(1):942–951, 2011.

[14] Mieczysław A Kłopotek. An aposteriorical clusterability criterion for k-means++ and simplicity of clustering. *SN Computer Science*, 1(2):1–38, 2020.

[15] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, Cambridge, 2008.

[16] Leonard Kaufman and Peter J Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. John Wiley & Sons, 2009.

[17] Dheeraj Kumar and James C Bezdek. Visual approaches for exploratory data analysis: A survey of the visual assessment of clustering tendency (vat) family of algorithms. *IEEE Systems, Man, and Cybernetics Magazine*, 6(2):10–48, 2020.

[18] G.R. Cross and A.K. Jain. Measurement of clustering tendency. In A.K. MAHALANABIS, editor, *Theory and Application of Digital Control*, pages 315–320. Pergamon, New Dehli, India, 1982.

[19] Richard G Lawson and Peter C Jurs. New index for clustering tendency and its application to chemical problems. *Journal of chemical information and computer sciences*, 30(1):36–41, 1990.

[20] Guangzhou Zeng and Richard C Dubes. A comparison of tests for randomness. *Pattern recognition*, 18(2):191–198, 1985.

[21] Richard C Dubes and Guangzhou Zeng. A test for spatial homogeneity in cluster analysis. *Journal of classification*, 4(1):33–56, 1987.

[22] Petros Drineas, Alan Frieze, Ravi Kannan, Santosh Vempala, and V Vinay. Clustering large graphs via the singular value decomposition. *Machine learning*, 56(1):9–33, 2004.

[23] Amit Banerjee and Rajesh N Dave. Validating clusters using the hopkins statistic. *International conference on fuzzy systems*, 1:149–153, 2004.

[24] William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

[25] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[26] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, 2005(09):P09008, 2005.

[27] Joseph C Dunn. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1):95–104, 1974.

[28] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[29] Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

[30] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, 28(1):100–108, 1979.

[31] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231, Portland, Oregon, 1996. AAAI Press.

[32] Fan RK Chung and Fan Chung Graham. *Spectral graph theory*. American Mathematical Soc., Providence, RI, 1997.

[33] Jeffrey D Banfield and Adrian E Raftery. Model-based gaussian and non-gaussian clustering. *Biometrics*, 49:803–821, 1993.

[34] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[35] Rainer Burkard, Mauro Dell'Amico, and Silvano Martello. *Assignment problems: revised reprint*. SIAM - Society of Industrial and Applied Mathematics, Philadelphia, 2012.

[36] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[37] Fevzi Alimoglu and Ethem Alpaydin. Combining multiple representations for pen-based handwritten digit recognition. *Turkish Journal of Electrical Engineering and Computer Sciences*, 9:1–12, 2001.

[38] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[39] Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[40] Stefano Monti, Pablo Tamayo, Jill Mesirov, and Todd Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1):91–118, 2003.

[41] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *ICISSp*, 1:108–116, 2018.

[42] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[43] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[44] Aristides Gionis, Heikki Mannila, and Panayiotis Tsaparas. Clustering aggregation. *Acm transactions on knowledge discovery from data (tkdd)*, 1(1):4–es, 2007.

[45] Martin Kleppmann. *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems.* " O'Reilly Media, Inc.", 2017.

[46] Zhou Tong, Xin Yuan, Scott Pakin, and Michael Lang. Performance and accuracy trade-offs of hpc application modeling and simulation. In *IEEE International Parallel and Distributed Processing Symposium*, pages 774–783, 05 2018.

[47] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Alec F. Diallo is currently a Ph.D. student at the University of Edinburgh. He received a joint Integrated Master's degree from Mundiapolis University and ESIEE Paris, with a focus on Computer Science and Electrical Engineering. His current research seeks to bridge the gap between the ever-evolving nature of cyber threats and the security and privacy of users' data on networked systems, by using Artificial Intelligence to build automatic threat detection and counteraction mechanisms.

Paul Patras is an Associate Professor in the School of Informatics at the University of Edinburgh, where he leads the Mobile Intelligence Lab – a multi-disciplinary team that pursues research at the intersection of network engineering and artificial intelligence, to improve the analysis, resilience, and management of next generation mobile systems. He is also a co-founder and CEO of Net AI, a pioneering university spinout specializing in AI-driven network analytics. He has served on the organizing committee on several conferences and workshops in his field, and advised the ITU-T Focus Group on Machine Learning for Future Networks including 5G. Paul holds M.Sc. and Ph.D. degrees from Universidad Carlos III de Madrid (UC3M) and he was the recipient of a prestigious Chancellor's Fellowship awarded by the University of Edinburgh.