Supplementary Information
A Fast and More Accurate Seed-and-Extension Density-based Clustering Algorithm

Ming-Hao Tung, Yi-Ping Phoebe Chen Senior Member, IEEE,
Chen-Yu Liu and Chung-Shou Liao Senior Member, IEEE

Figure S1 shows an example illustrating the DP algorithm: Figure 1(left) presents the distribution of an input set of data points, and Figure 1(middle) shows the corresponding decision graph of the dataset. In this example, there are three significant outliers: A, B and C. Point A has the highest density over the other data points. We call this data point an *absolute density peak*. Based on the assumptions of DP, these three outliers are suggested to be the cluster centers. In the procedure of assigning data points (Subroutine 3), each remaining data point *i* is assigned to the cluster in which *i*'s closest data point with a higher density (denoted by CHD_i) lies. The assignment process is illustrated in Figure 1(Right), where the height of each point i indicates its density ρ_i and the width indicates its distance. As shown in Figure 1(right), the density peaks are distinguished from other points because points A, B and C have the highest local density ρ and a large distance δ between each other. Hence, A, B and C represent the centers in the red, green and blue clusters, respectively in Figure 1(Left).

We remark that in some datasets, the number of clusters, *k*, cannot be simply determined. Moreover, even given the number of clusters, *k* for an input dataset, one cannot easily use its decision graph to decide what data points should be chosen as cluster centers.

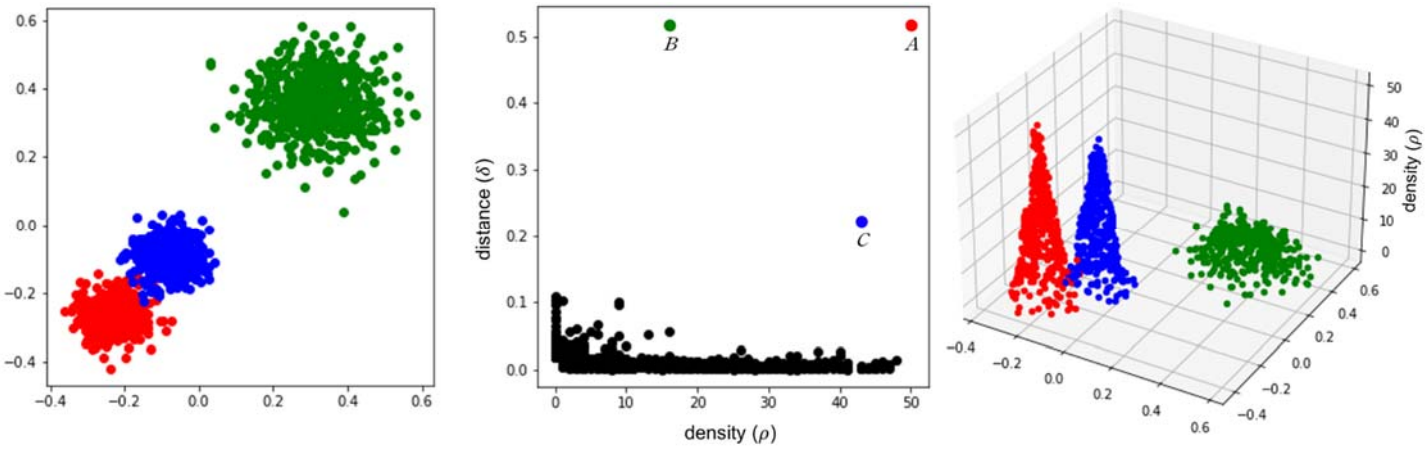

Figure S1. (Left) The distribution of a set of data points in the two-dimensional space; (Middle) The corresponding decision graph of the dataset, where the points A, B and C represent the centers of the red, green and blue clusters, respectively in Fig. 1 (Left); (Right) The concept of the cluster assignment from each remaining point to the three clusters by linking its CHD

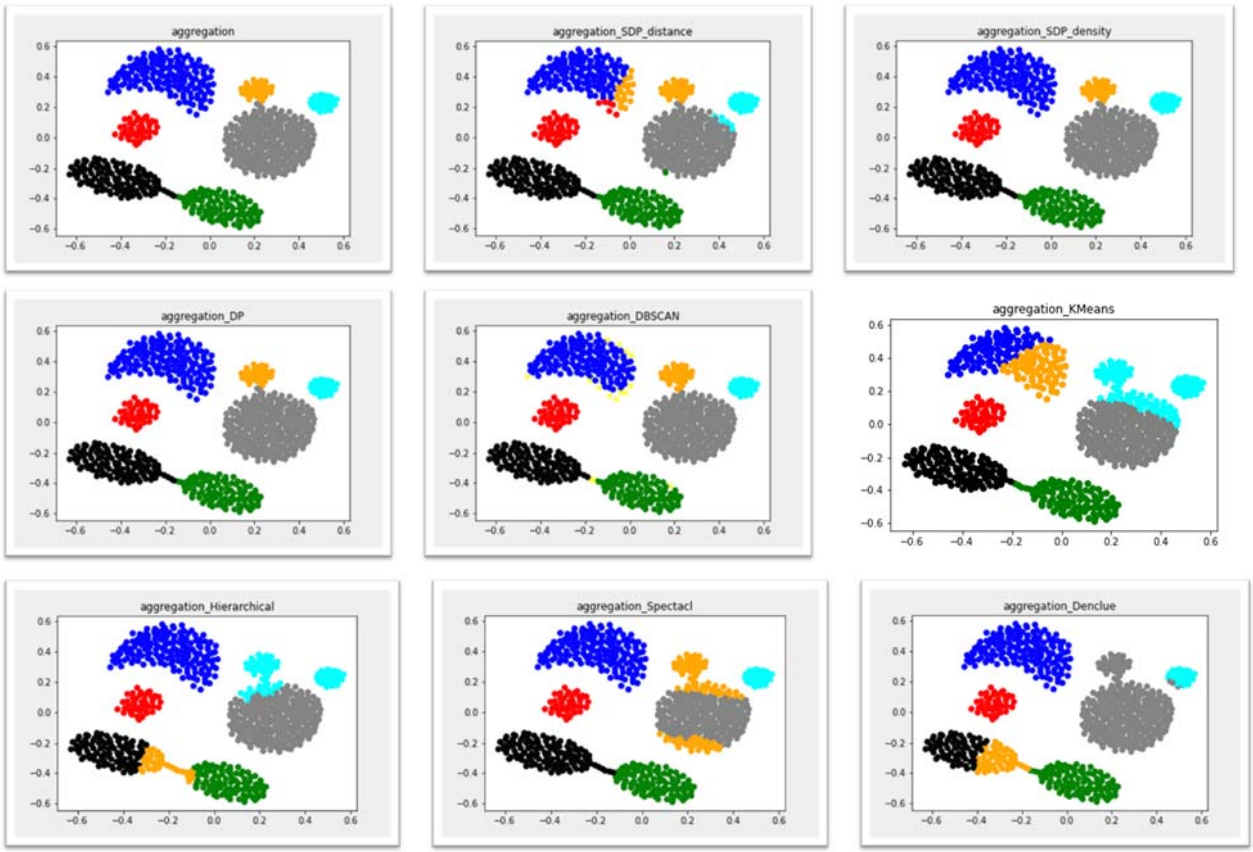

Figure S2. Illustration of the difference between the clustering result of SDPdist, SDPden, DP, DBSCAN, K-Means, Hierarchical, SPECTACL, DENCLUE 2.0 for the aggregation dataset

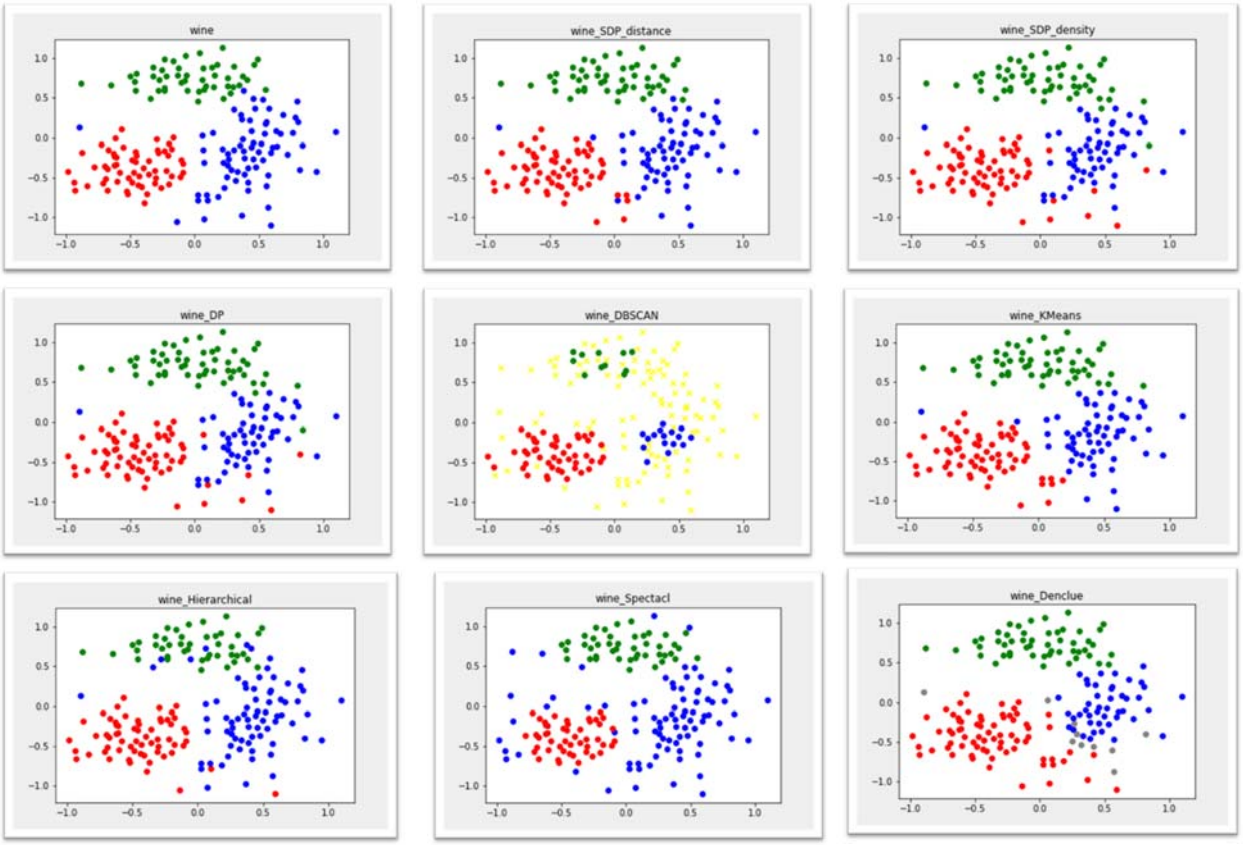

Figure S3. Illustration of the difference between the clustering result of SDPdist, SDPden, DP, DBSCAN, K-Means, Hierarchical, SPECTACL, DENCLUE 2.0 for the wine dataset

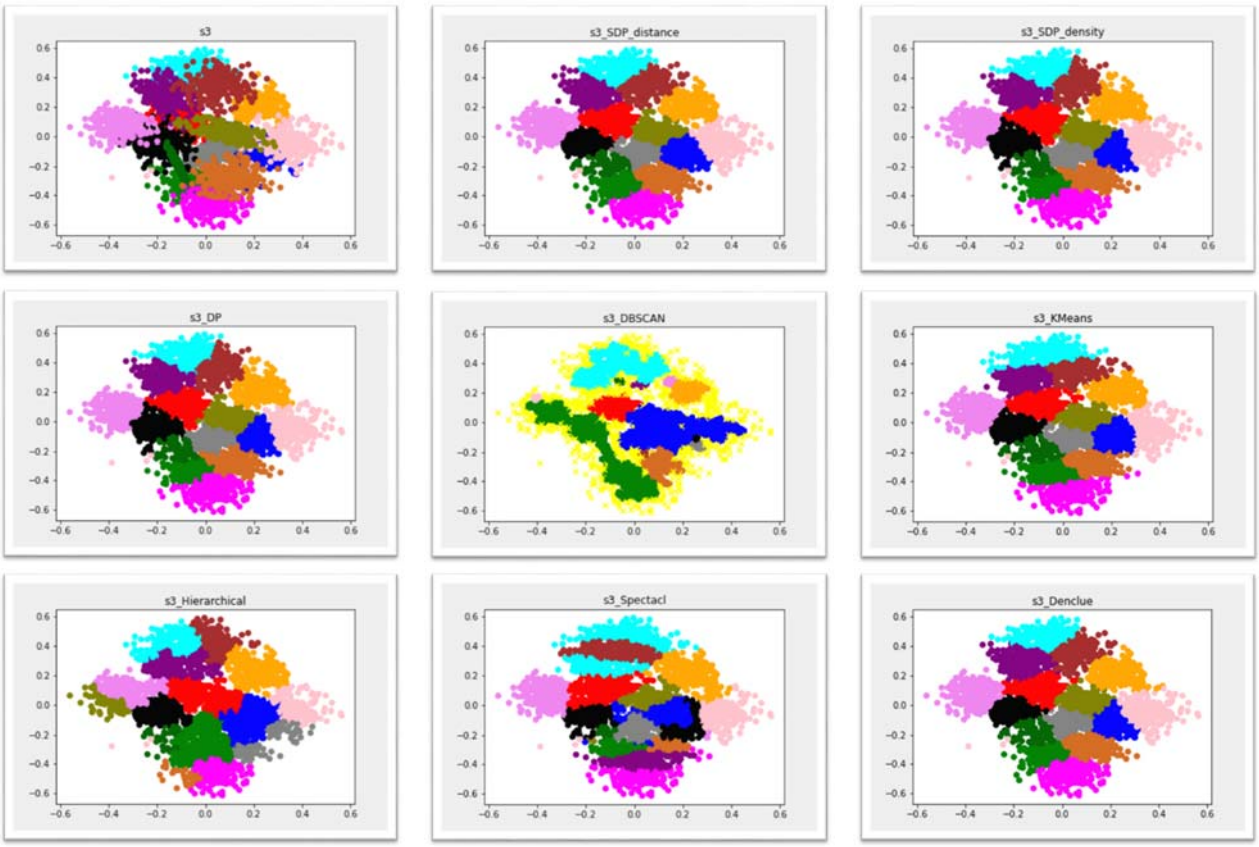

Figure S4. Illustration of the difference between the clustering result of SDPdist, SDPden, DP, DBSCAN, K-Means, Hierarchical, SPECTACL, DENCLUE 2.0 for the s3 dataset

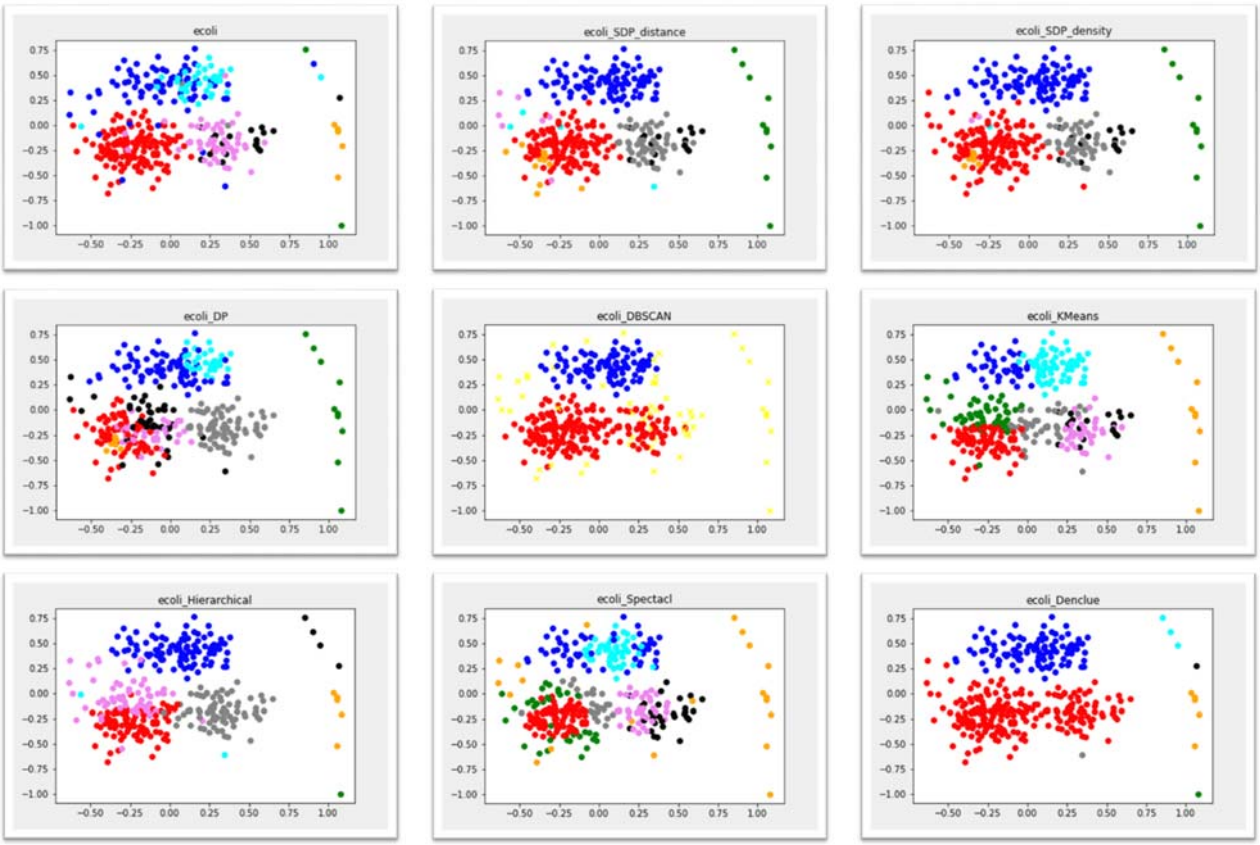

Figure S5. Illustration of the difference between the clustering result of SDPdist, SDPden, DP, DBSCAN, K-Means, Hierarchical, SPECTACL, DENCLUE 2.0 for the ecoli dataset

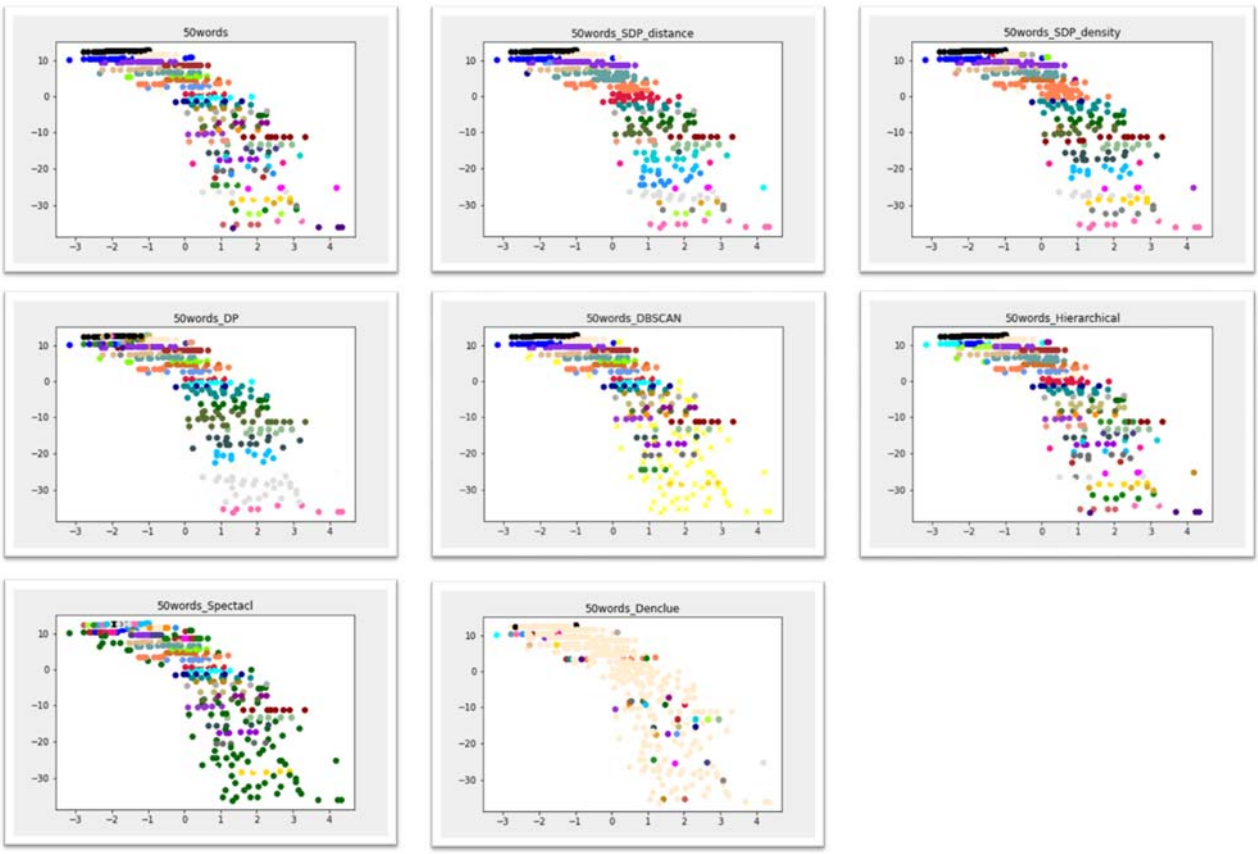

Figure S6. Illustration of the difference between the clustering result of SDPdist, SDPden, DP, DBSCAN, Hierarchical, SPECTACL, DENCLUE 2.0 for the 50 words dataset

From Figures S1 to S6, we use Shepard diagrams with multidimensional scaling (MDS) which can provide a visual representation of the pattern of proximities among a set of objects/data points, to illustrate the superior performance of our SDP algorithm in a clearer way. Note that MDS can be used to interpret “pairwise similarity/distance” between *n* input objects into a configuration of *n* points mapped to a two-dimensional space, i.e., a Shepard diagram. More precisely, a Shepard diagram is a scatterplot in which the X-axis usually corresponds to the input proximities and the Y-axis corresponds to the MDS distances. However, the axes actually are, in themselves, meaningless, and moreover, the orientation of the diagram is arbitrary. Shepard diagrams with MDS just indicate how close a point is to other points.