APPENDIX

We provide supplementary material, from two sets of experiments, to show different aspects of the SD-DP algorithm. The first set contains confirmative experiments on datasets with known ground truth. We show in particular: A. the merge structure on the MNIST dataset [13], and B. the comparison in accuracy between DP and SD-DP on five synthetic datasets [8]. DP failed on one of them. The second set contains exploratory experiments on data to be categorized. We show: C. the conditioning effect of splits on cluster merges, with dataset PBMCs-8k [36], and D. new data links captured by SD-DP on `GloVe` word vectors [21].

A. *Hierarchical merge*

We present the hierarchical structure of the *unsupervised* merging results on the MNIST dataset. SD-DP achieves superior clustering accuracy, compared to the *supervised* merging with the modified DP, as summarized in Table 1. At $k = 48$, SD-DP identifies 53 local maxima, which constitute the initial configuration. Clustering is hierarchical in nature; the local maxima provide a configuration at a high level in the hierarchy. The initial clusters are automatically, hierarchically merged into 11 clusters, with threshold $\tau = 0.12$ in (7). Fig. 7 and 8 display the merge hierarchy in two ways: i) a two-level partition of the kNN matrix $\mathbf{G}_k$: a fine 53×53 partition (blue blocks) corresponding to the initial configuration and a coarse 11×11 partition (red, dashed lines) corresponding to the 11 merged clusters, and ii) a dendrogram displaying the merge dependence, with the local maxima at the base level. By adapting the threshold, the two clusters associated with digit 1 are merged. In addition, up to $2\,\%$ improvement in the F_1 scores is achieved with cluster splits.

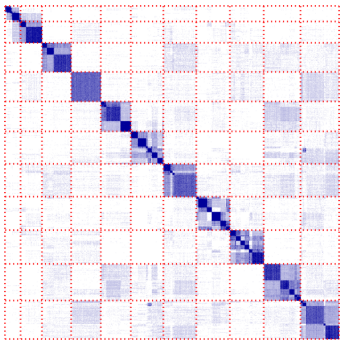


Fig. 7: kNN matrix $\mathbf{G}_k$ of MNIST handwritten digits, in a two-level partition. A fine 53×53 partition (blue blocks) corresponds to the initial configuration and a coarse 11×11 partition (red, dashed lines) corresponds to 11 merged clusters.

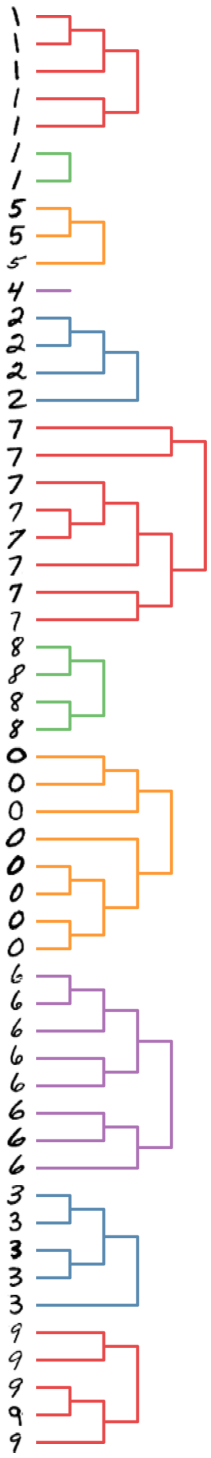


Fig. 8: Unsupervised hierarchical merging of 53 local maxima ($k = 48$) on MNIST handwritten digits. The dendrogram depicts merge dependence, with the local maxima at the base level.

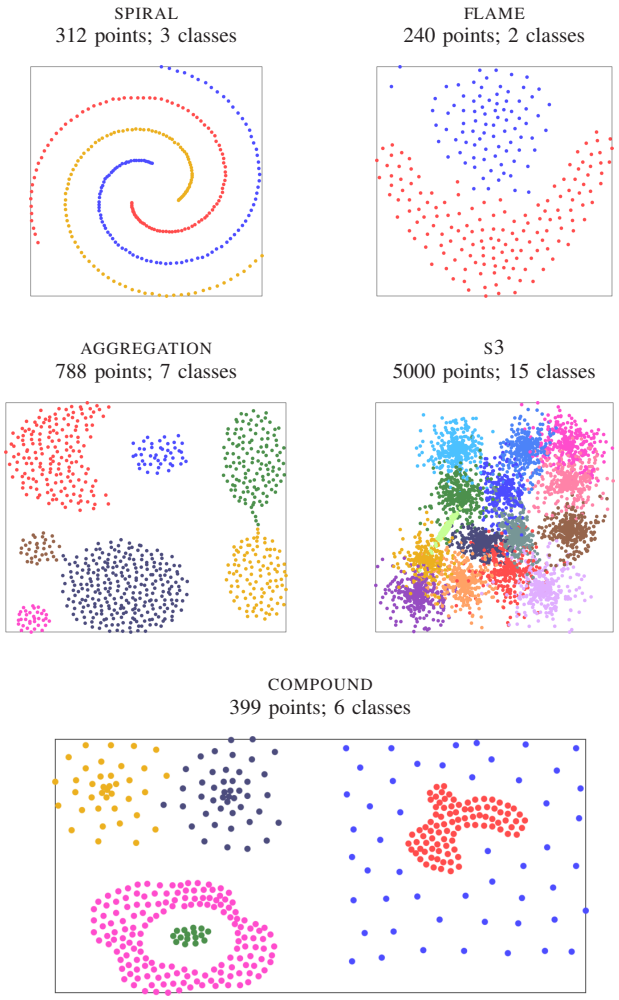


Fig. 9: Color-coded ground truth for each of the five synthetic datasets used to compare DP and SD-DP. The subtitle over each dataset specifies the name, the number of data points, and the number of true classes.

B. Comparison in accuracy between SD-DP and DP

We provide comparisons in accuracy between SD-DP and DP on five synthetic two-dimensional datasets, which are publicly available [8]. These datasets are often used for benchmarking model-based or non-parametric clustering algorithms. See Table 2 for a data summary and Fig. 9 for the ground truth. The parameters used for both DP and SD-DP are also listed in Table 2.

TABLE 2: Summary of datasets and algorithm parameters. **Left column:** datasets used for the experiments, each with N points in L clusters; **Right columns:** Algorithm parameters for DP and SD-DP. For DP, d_{H} is hard cut-off and d_{S} is soft cut-off.

Dataset			**DP**		**SD-DP**
Name	N	L	d_{H}	d_{S}	k
SPIRAL	312	3	2.30	3.10	7.00
FLAME	240	2	1.24	1.60	7.00
AGGREGATION	788	7	2.67	1.80	23.00
S3	5000	15	41,884	41,884	133.00
COMPOUND	399	6	3.70	3.10	9.00

The first four datasets were used in the original DP paper [22], but no specific parameter values were provided. We also include the COMPOUND dataset [33], see the dataset at the bottom of Fig. 9. This dataset contains two interesting mixed structures, we refer to them as mixtures A and B. Mixture A, located in the right part, is composed of two modes with different densities but with overlapping support. Both DP and SD-DP separate this mixture from the rest correctly, but they are unable to decouple the modes within the mixture. The reason is simple: the DP principle assumes spatial separation of cluster centers. This is not necessarily a fault of the DP assumptions. Once such a mixture is identified, one may make further inquiry and analysis of the mixture. Mixture B, located in the bottom left part, has two modes of non-overlapping support, but the support of one is surrounded by the support of the other. Mixture B is not excluded by the DP principles. SD-DP renders Mixture B correctly, but DP fails.

We show density histogram and decision graph for each algorithm on every dataset. By Theorem 1, the critical curve $\gamma^* = 1$ on the SD-DP decision graph separates the local maxima from the rest. By Theorem 3, on the DP decision graph the local maxima are above the horizontal line $\delta = d_c$. The comparisons are shown in Fig. 12 to 16. For DP we use the best parameters we can find.

C. Splitting effect on cluster merges

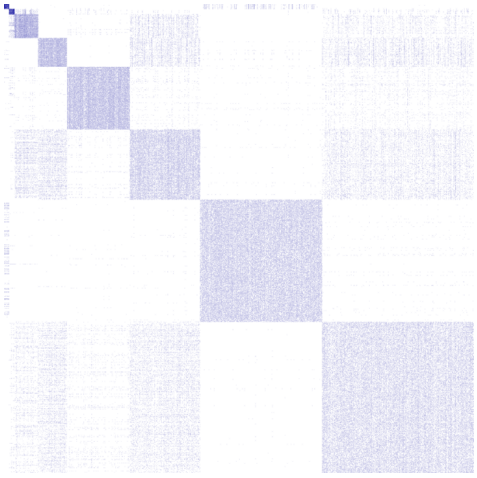
In Fig. 10, we use the dataset PBMCs-8k to show: i) the cluster configuration by the local maxima, ii) the subclusters that were split from each initial cluster centered at a local maximum, and iii) the merged cluster configuration. The result with split conditioning is substantially better than without.

Fig. 10a shows the block partition of the kNN matrix $\mathbf{G}_k$ by the local maxima configuration. Fig. 10b shows the cluster configuration modified by splits. Fig. 10c shows the merge after splitting. The off-diagonal interaction strength h is reduced by $2.77\times$, while the area f of the diagonal blocks is only increased by $9\,\%$ than the initial configuration.

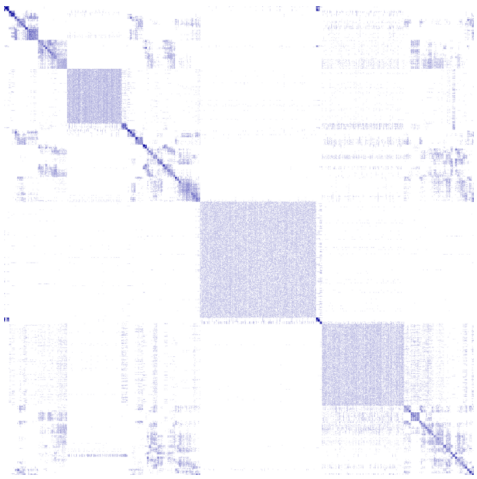
D. Discovery of data links

We use the `GloVe` [21] word vectors of 300 dimensions from `Wikipedia 2014 + Gigaword 5`.[5] The vocabulary data consists of 400,000 words. The kNN graph with $k = 5$ was computed using Euclidean distance.

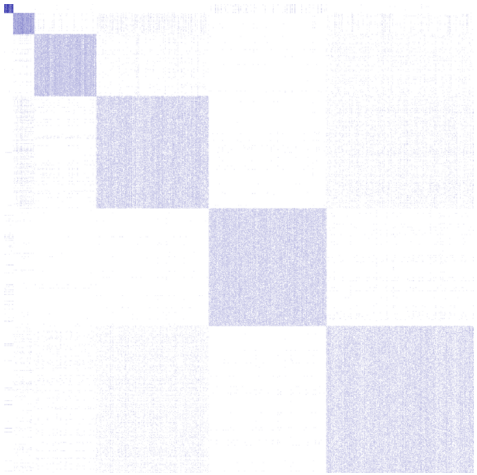
The dual local densities are related to the word co-occurrence frequencies. Intrinsic structures of the words are revealed by the SD-DP ascending trees. The trees depict the statistical hierarchy of the word semantics; a word with higher density has more general meaning, while a word with lower density has more specific context. In Fig. 11 we show the upper levels of three ascending trees rooted at the local maxima `movie`, `merlot`, and `physicians`. The structures show an interesting, original way to search and retrieve words, in depth and breadth, simultaneously.



(a) Initial cluster configuration by local maxima.



(b) Clusters split into sub-clusters.



(c) Final cluster configuration after merging.

Fig. 10: Inspection of SD-DP split-and-merge results on the dataset PBMCs-8k. The kNN matrix $\mathbf{G}_k$ ($k = 35$) is shown at each step.

[5] https://nlp.stanford.edu/projects/glove/

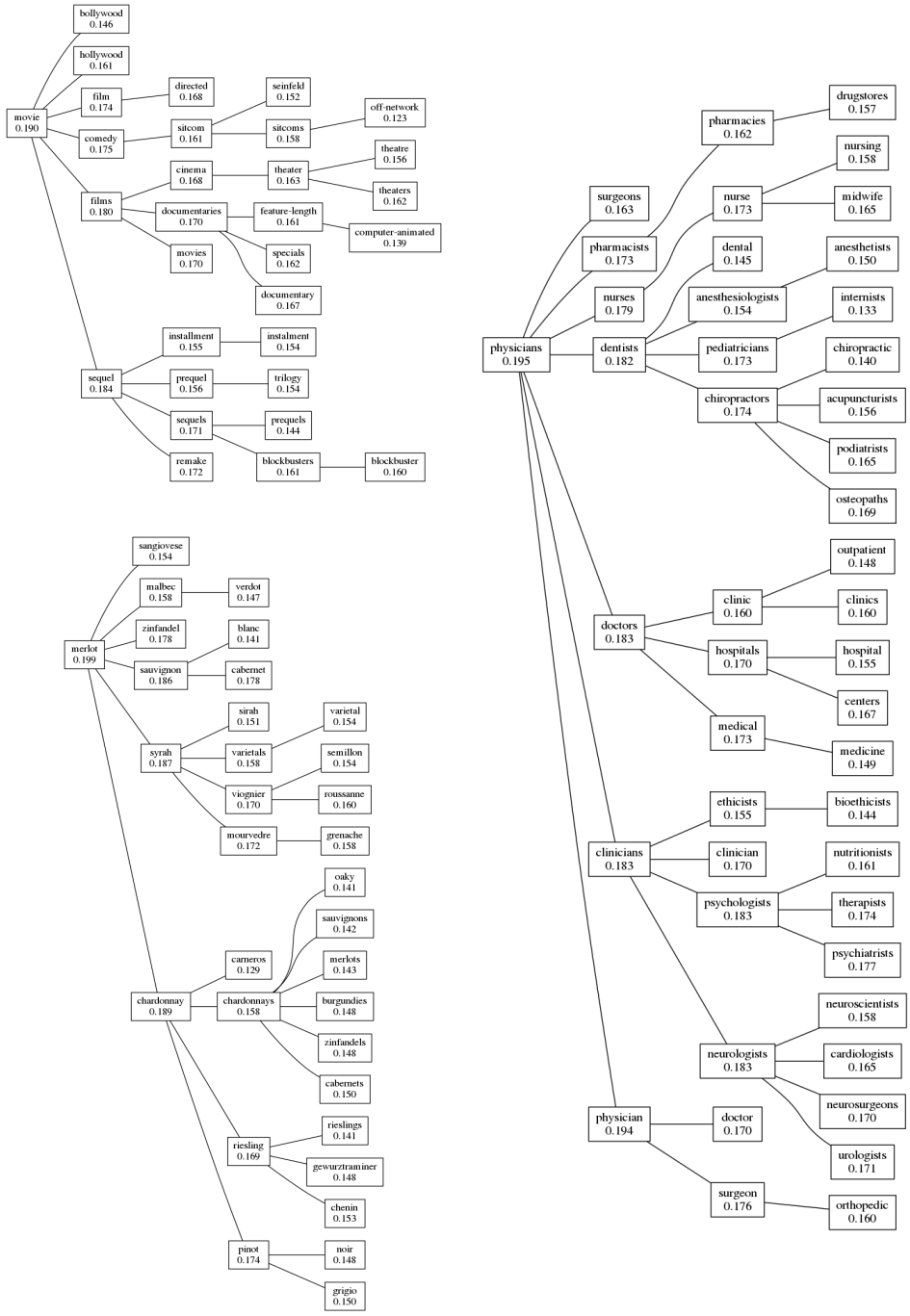
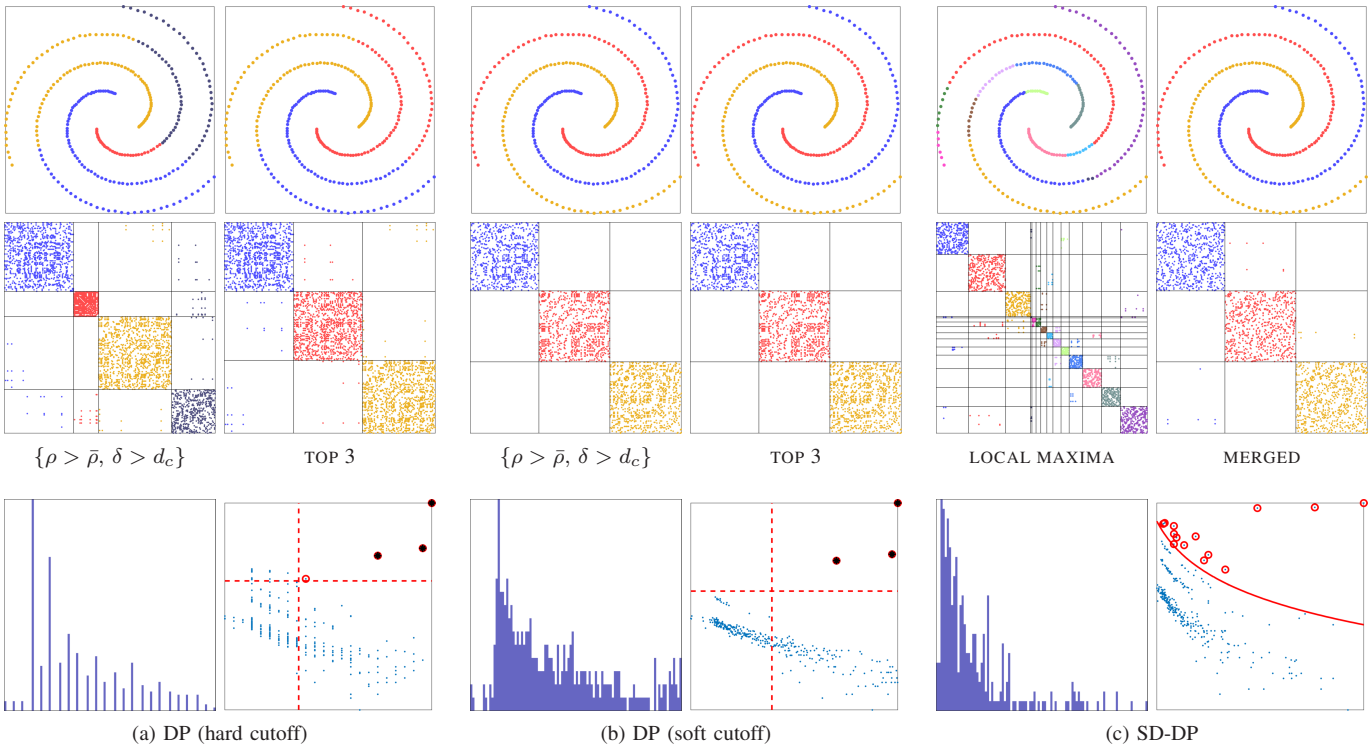
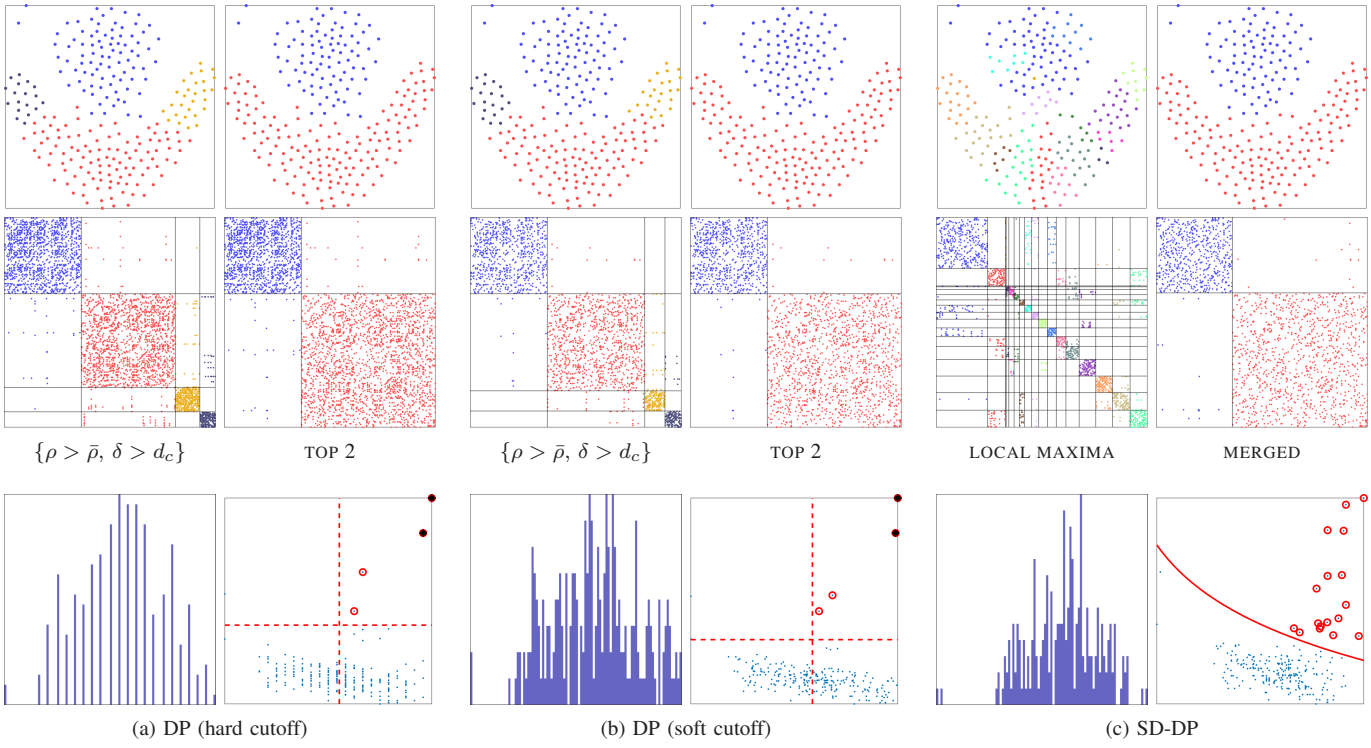


Fig. 11: The upper levels of three SD-DP ascending trees rooted at the local maxima `movie`, `merlot`, and `physicians`. The dual local densities are related to the word co-occurrence frequencies. Each word is annotated by its dual local density.



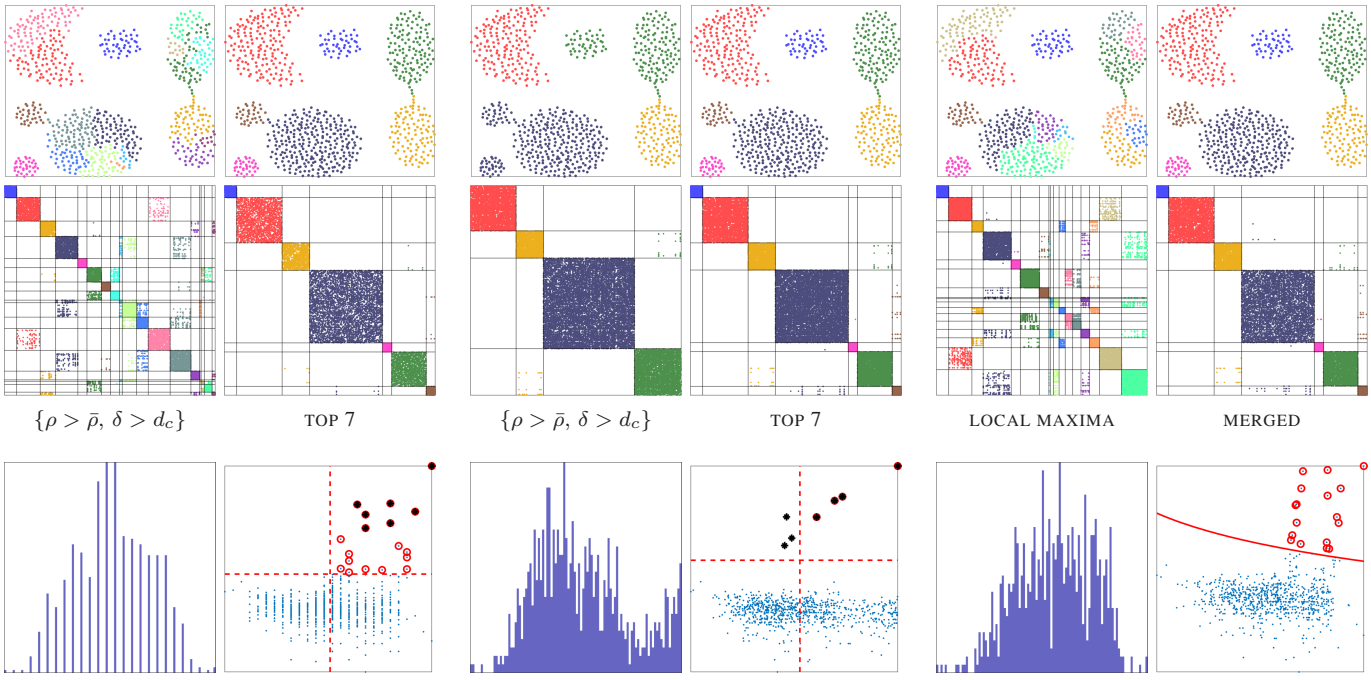
(a) DP (hard cutoff) (b) DP (soft cutoff) (c) SD-DP

Fig. 12: Comparison between DP (soft and hard threshold) and SD-DP cluster configurations on dataset SPIRAL. **Top row**: Estimated cluster labels, color-coded. **Second row**: Block-per-cluster near-neighbor graph matrix. rNN graphs are shown in DP and kNN graphs in SD-DP. **Third row**: Density histograms and decision graphs.



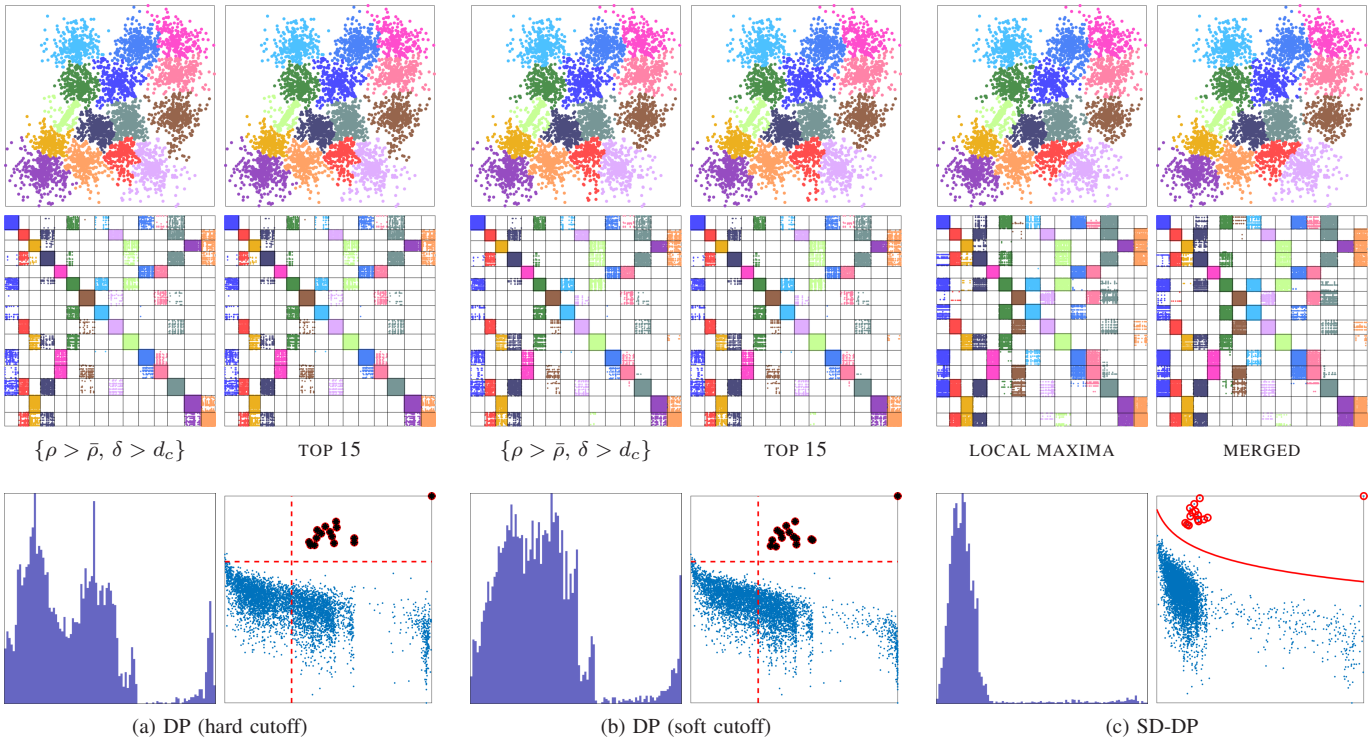
(a) DP (hard cutoff) (b) DP (soft cutoff) (c) SD-DP

Fig. 13: Comparison between DP (soft and hard threshold) and SD-DP cluster configurations on dataset FLAME. **Top row**: Estimated cluster labels, color-coded. **Second row**: Block-per-cluster near-neighbor graph matrix. rNN graphs are shown in DP and kNN graphs in SD-DP. **Third row**: Density histograms and decision graphs.



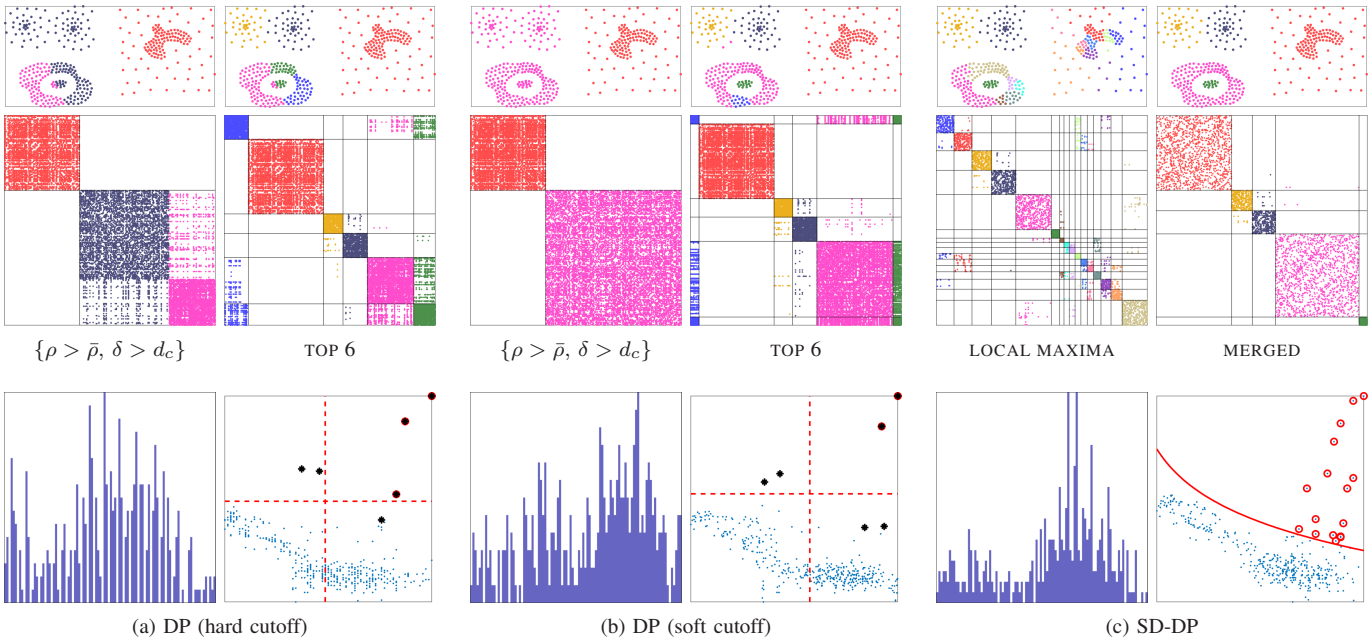
(a) DP (hard cutoff) (b) DP (soft cutoff) (c) SD-DP

Fig. 14: Comparison between DP (soft and hard threshold) and SD-DP cluster configurations on dataset AGGREGATION. **Top row**: Estimated cluster labels, color-coded. **Second row**: Block-per-cluster near-neighbor graph matrix. rNN graphs are shown in DP and kNN graphs in SD-DP. **Third row**: Density histograms and decision graphs.



(a) DP (hard cutoff) (b) DP (soft cutoff) (c) SD-DP

Fig. 15: Comparison between DP (soft and hard threshold) and SD-DP cluster configurations on dataset S3. **Top row**: Estimated cluster labels, color-coded. **Second row**: Block-per-cluster near-neighbor graph matrix. rNN graphs are shown in DP and kNN graphs in SD-DP. **Third row**: Density histograms and decision graphs.



(a) DP (hard cutoff) (b) DP (soft cutoff) (c) SD-DP

Fig. 16: Comparison between DP (soft and hard threshold) and SD-DP cluster configurations on dataset COMPOUND. **Top row**: Estimated cluster labels, color-coded. **Second row**: Block-per-cluster near-neighbor graph matrix. rNN graphs are shown in DP and kNN graphs in SD-DP. **Third row**: Density histograms and decision graphs. SD-DP succeeds in clustering correctly the data points at the bottom left region.