# Experiments on Density-Constrained Graph Clustering* **

Robert Görke, Andrea Schumm, and Dorothea Wagner

Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Germany

**Abstract.** Clustering a graph means identifying internally dense subgraphs which are only sparsely interconnected. Formalizations of this notion lead to measures that quantify the quality of a clustering and to algorithms that actually find clusterings. Since, most generally, corresponding optimization problems are hard, heuristic clustering algorithms are used in practice, or other approaches which are not based on an objective function. In this work we conduct a comprehensive experimental evaluation of the qualitative behavior of greedy bottom-up heuristics driven by cut-based objectives and constrained by intracluster density, using both real-world data and artificial instances. Our study documents that a greedy strategy based on local movement is superior to one based on merging. We further reveal that the former approach generally outperforms alternative setups and reference algorithms from the literature in terms of its own objective, while a modularity-based algorithm competes surprisingly well. Finally, we exhibit which combinations of cut-based inter- and intracluster measures are suitable for identifying a hidden reference clustering in synthetic random graphs.

## 1 Introduction

Graph clustering aims at finding subsets of vertices that are densely connected with each other but sparsely connected with the remainder of the graph. In the last decades, interest in graph clustering algorithms has grown rapidly, with applications ranging from customer recommendation systems to the analysis of networks describing social ties or protein-protein interaction. A variety of measures have been proposed, which are used to assess and compare different clusterings and to guide the design of algorithms. Traditional methods from algorithmics often focus on sparse cuts with respect to measures like conductance [18] or expansion [16], while, independent from that, a measure called modularity [21] proved to yield meaningful clusterings on a wide range of application data.

Recently, we systematically assembled a range of self-evident intracluster density and intercluster sparsity measures for clusterings, where the latter are based on conductance , expansion and density of the cuts induced by the clusters [14].

* Published in the Proceedings of the 14th Meeting on Algorithm Engineering and Experiments (ALENEX'12). SIAM, 2012

** This work was partially supported by the DFG under grant WA 654/19-1

We further formally stated the problem Density-Constrained Clustering (DCC), where the objective is to optimize intercluster sparsity with the constraint that the intracluster density must exceed a given threshold. As optimal polynomial-time algorithms for DCC are unknown, we investigated how different combinations of intracluster sparsity and intercluster density measure influence the efficiency of a greedy optimization strategy based on cluster merging. However, little is known about its qualitative behavior in practical scenarios, and an experimental evaluation of DCC has yet been missing.

**Our Contribution.** We provide a comprehensive study of the practical behavior of greedy graph clustering heuristics driven by cut-based objectives and constrained by intracluster density. We give evidence that, in general, greedy algorithms based on local vertex moves lead to better quality than the corresponding merge-based algorithm. We then compare the move-based algorithm to a set of reference algorithms from the literature, both with respect to the objective of DCC and their ability to reconstruct planted partitions in a family of synthetic graphs. We find that the greedy move algorithm compares favorably to most reference algorithms in the context of DCC, while a comparison with the modularity-based algorithm shows that optimizing modularity implicitly yields good results for some variants of DCC. Experiments with planted partition graphs suggest that certain combinations of inter- and intracluster measures are effective in finding the hidden clustering, while others clearly fail. Together with observations about the number of identified clusters, this yields valuable insights about the behavior of the respective intra- and intercluster density measures.

**Related Work.** Related clustering algorithms are Iterative Conductance Cutting [18], Markov-Clustering [10], Geometric MST Clustering [6] and a modularity-based greedy algorithm based on vertex moves [22]; we use these as reference algorithms. Kannan et al. propose to minimize the cut between clusters, subject to a guaranteed intracluster conductance [18], which is closely related to DCC. They further show that Iterative Conductance Cutting has polylogarithmic approximation guarantees on both of these measures. Brandes et al. conduct an experimental study on the performance of Iterative Conductance Cutting, Markov-Clustering and Geometric MST Clustering, both with respect to quality and running times [7]. A similar, but more recent study can be found in [19]. Flake et al. give a clustering algorithm with provable, but interdependent bounds on both intra- and a variant of intercluster expansion. The notion of modularity was introduced in [21], an extensive and recent overview of the research on it can be found in [12]. Apart from these, there is a huge number of publications on graph clustering, for an overview see [17, 4].

## 2 Preliminaries

**Notation.** Let $G = (V, E)$ be an undirected, unweighted, and simple graph, i.e. $G$ is loopless and has no parallel edges. In the following, $n$ will always denote the number of vertices and $m$ the number of edges in $G$.

Table 1: Density measures

| intracluster density | | |
|---|---|---|
| global | gid | $\frac{\sum_{C\in\mathcal{C}} m_C}{\sum_{C\in\mathcal{C}} \binom{n_C}{2}}$ |
| minimum | mid | $\min_{C\in\mathcal{C}} \frac{m_C}{\binom{n_C}{2}}$ |
| average | aid | $\frac{1}{\lvert\mathcal{C}\rvert} \sum_{C\in\mathcal{C}} \frac{m_C}{\binom{n_C}{2}}$ |

| intercluster density | | |
|---|---|---|
| global | gxd | $\frac{\sum_{A\neq B\in\mathcal{C}} m_{A,B}}{\sum_{A\neq B\in\mathcal{C}} n_A n_B}$ |
| maximum | mixd | $\max_{C\in\mathcal{C}} \frac{x_C}{n_C n_{V\setminus C}}$ |
| average | aixd | $\frac{1}{\lvert\mathcal{C}\rvert} \sum_{C\in\mathcal{C}} \frac{x_C}{n_C n_{V\setminus C}}$ |

| intercluster conductance | | |
|---|---|---|
| maximum | mixc | $\max_{C\in\mathcal{C}} \frac{m_{C,V\setminus C}}{\min\{v_C, v_{V\setminus C}\}}$ |
| average | aixc | $\frac{1}{\lvert\mathcal{C}\rvert} \sum_{C\in\mathcal{C}} \frac{m_{C,V\setminus C}}{\min\{v_C, v_{V\setminus C}\}}$ |

| intercluster expansion | | |
|---|---|---|
| maximum | mixe | $\max_{C\in\mathcal{C}} \frac{m_{C,V\setminus C}}{\min\{n_C, n_{V\setminus C}\}}$ |
| average | aixe | $\frac{1}{\lvert\mathcal{C}\rvert} \sum_{C\in\mathcal{C}} \frac{m_{C,V\setminus C}}{\min\{n_C, n_{V\setminus C}\}}$ |

| intercluster edges | | |
|---|---|---|
| global | nxe | $\sum_{A\neq B\in\mathcal{C}} m_{A,B}$ |

| modularity | | |
|---|---|---|
| global | mod | $\frac{\sum_{C\in\mathcal{C}} m_C}{m} - \frac{\sum_{C\in\mathcal{C}} v_C^2}{4m^2}$ |

For two subsets $A$ and $B$ of $V$, $m_{A,B} := |\{\{u,v\} \in E \mid u \in A, v \in B\}|$ is the number of edges between $A$ and $B$, $n_A := |A|$ is the number of vertices in $A$, $m_A := |E(A)|$ is its number of intracluster edges and $x_A := m_{A,V\setminus A}$ the number of intercluster edges incident to $A$. Further, the *volume* $v_A$ of $A$ is defined as $v_A := \sum_{v\in A} \deg(v)$. The *conductance* of a cut $(S,T)$ measures the bottleneck between $S$ and $T$, defined as $\frac{m_{S,T}}{\min\{v_S,v_T\}}$; *expansion* substitutes volume by cardinality: $\frac{m_{S,T}}{\min\{n_S,n_T\}}$. The *density* (or *sparsity*) of a cut is $\frac{m_{S,T}}{n_S n_T}$, which equals the *uniform minimum-ratio cut*. We restrict ourselves to disjoint clusters in this work, this means, if $\mathcal{C} = \{C_1, \ldots, C_k\}$ is a partition of $V$, we call $\mathcal{C}$ a *clustering* of $G$ and the sets $C_i$ *clusters*. The cluster containing vertex $v$ is $\mathcal{C}(v)$ and the clustering that results from moving vertex $v$ to cluster $D$, i.e. $\big(\mathcal{C}\setminus\{C(v),D\}\big)\cup\{C(v)\setminus v, D\cup\{v\}\}$, is abbreviated by $\mathcal{C}_{v\to D}$. A clustering is *trivial* if either $k = 1$ (*all-clustering*), or each cluster contains only one element (*singletons*). We identify a cluster $C$ with the set of nodes it constitutes and with its vertex-induced subgraph of $G$. Then $E(\mathcal{C}) := \bigcup_{C\in\mathcal{C}} E(C)$ are called *intracluster* edges and $E \setminus E(\mathcal{C})$ *intercluster* edges. A *clustering measure* is a function that maps clusterings to real numbers, thereby assessing the quality of a clustering. We define high quality to correspond to high (low) values of intracluster (intercluster) measures and will always denote intracluster density measures with $i$ and intercluster density measures with $x$, unless otherwise stated.

**Intracluster Density and Intercluster Sparsity Measures.** All intercluster measures we use are based on *cuts* or *$k$-way cuts*. Separating a single cluster from the remaining vertices induces a cut, whose sparsity can be evaluated using density, conductance or expansion. This defines a set of sparsity values for the whole clustering, from which we can either compute the average or the maximum, yielding *maximum/average intercluster density/conductance/expansion* (mixd, aixd, mixc, aixc, mixe and aixe)[1]. Another point of view is to evaluate the clustering as a whole, i.e. to assess the sparsity of the induced $k$-way cut

[1] Note that we keep the $i$ in the abbreviations, although in contrast to [14], we do not distinguish between pairwise and isolated measures

directly. We do this by either counting the number of intercluster edges (nxe) or by dividing the number of intercluster edges by the maximum possible number, i.e. the number of intercluster pairs (gxd). It is possible to use similar, cut-based measures for intracluster density. However, even evaluating these measures for a given clustering is $NP$-hard, such that clustering algorithms usually work with approximations or bounds [18, 11, 7]. As we intend to use intracluster density measures as constraints in greedy bottom-up algorithms, it is crucial to be able to evaluate them efficiently. We therefore use a more practical approach and define intracluster density as the ratio of the number of intracluster edges and the number of intracluster pairs. Evaluating this globally leads to *global intracluster density* (gid), whereas the average and minimum of all clusters yields *average* and *minimum intracluster density* (aid and mid).

Table 1 summarizes the formalizations of all measures considered. Note that, in contrast to the set of measures used in [14], we omit the notions of *pairwise* densities as they turned out to be very prone to local minima if used with greedy bottom-up algorithms. Although it does not quite fit into this classification, Table 1 also includes the objective used by one of the reference algorithms, *modularity*, which simultaneously assesses intracluster density and intercluster sparsity by subtracting from the fraction of intracluster edges the expectation of this value in a random graph (high modularity corresponds to high quality).

**Density-Constrained Clustering.** Density-Constrained Clustering is the problem of optimizing intercluster density while retaining guarantees on the intracluster density. Considering each combination of intracluster and intercluster measure listed in Table 1 leads to a family of optimization problems. Slightly abusing the notation, we consider modularity as an intercluster density objective in this context.

*Problem 1 (*Density-Constrained Clustering(DDC)*).* Given a graph $G = (V, E)$, among all clusterings with an intracluster density of no less than $\alpha$, find a clustering $\mathcal{C}$ with optimum intercluster quality.

## 3 Greedy Algorithms for Density-Constrained Clustering

The following generic greedy algorithms heuristically minimize (maximize) the objective function of DCC for all density measures considered.

**Greedy Merge (GM).** Starting from singletons, the algorithm greedily merges pairs of clusters. In each step, among all pairs of clusters whose merge does not violate the constraint on the intracluster density, the merge with the largest benefit to the intercluster density is performed. We recently proposed this algorithm in the context of DDC [14] and classified combinations of intercluster and intracluster density with respect to the question how efficiently this algorithm can be implemented. Algorithms of these kind are common in the context of clustering point sets in $d$-dimensional space, where a basic constraint is that the number of clusters must not fall below a certain threshold. In the field of graph clustering, this algorithm is used to optimize modularity [8].

**Algo. 1:** Greedy Vertex Moving

**Input** : graph $G$, $\mathsf{inter}$, $\mathsf{intra}$, $\alpha$
**Output**: clustering $\mathcal{C}_0$ of $G$
$G^0 \leftarrow G$, $h \leftarrow 0$
**repeat**
  $\mathcal{C}^h \leftarrow \text{Singletons}(G^h)$
  $\mathcal{C}^h \leftarrow \text{LM}(G^h, C^h, \mathsf{intra}, \mathsf{inter}, \alpha)$
  $G^{h+1} \leftarrow \text{contract}(G^h, \mathcal{C}^h)$
  $h \leftarrow h + 1$
**until** *no more real contractions*
**while** $h \geq 0$ **do**
  $h \leftarrow h - 1$
  $\mathcal{C}^h \leftarrow \text{project}(\mathcal{C}^{h+1}, G^h)$
  $\mathcal{C}^h \leftarrow \text{LM}(G^h, \mathcal{C}^h, \mathsf{inter}, \mathsf{intra}, \alpha)$
**end**
**return** $\mathcal{C}^0$

**Algo. 2:** Local Moving (LM)

**Input** : graph $G$, clustering $\mathcal{C}_{\text{init}}$ of $G$, $\mathsf{inter}$, $\mathsf{intra}$, $\alpha$
**Output**: clustering $\mathcal{C}$ of $G$
$\mathcal{C} \leftarrow \mathcal{C}_{\text{init}}$
**repeat**
  **forall the** $v \in V$ **do**
    $\mathcal{A} \leftarrow \{C \in \mathcal{C} \mid \mathsf{intra}(\mathcal{C}_{v \to C}) \geq \alpha \text{ and } |E(v, C)| > 0\}$
    $N \leftarrow \arg \min_{C \in \mathcal{A} \cup \{\}} \{\mathsf{inter}(\mathcal{C}_{v \to C})\}$
    **if** $\mathsf{inter}(\mathcal{C}_{v \to N}) < \mathsf{inter}(\mathcal{C})$ **then**
      $\text{move}(v, N)$
    **end**
  **end**
**until** *no more changes*
**return** $\mathcal{C}$

**Greedy Vertex Moving (GVM).** The key ingredient of GVM (Algo. 1) is a subprocedure that tries to greedily improve the objective function by letting vertices move to neighboring clusters (Algo. 2). This subprocedure repeatedly iterates through the vertex set and, for each vertex, performs the most improving move (subject to the constraint), potentially isolating a vertex, or leaving it where it was, until a local optimum is reached. Starting with singletons, GVM first calls this subprocedure and contracts the resulting preliminary clustering into a super-graph, i.e. each cluster becomes a vertex weighted with the number of vertices it represents, and edges are summarized such that edge weights reflect the number of edges in the original graph. This whole process is iterated until local moving does not yield any further improvement, and results in a hierarchy of graphs with increasing coarseness. In the second phase (refinement), the hierarchy is unfurled step by step by projecting the clustering of the $i+1$-th level of the hierarchy to level $i$, i.e. the clusters in level $i$ are merged according to the clustering in level $i+1$. After each step, LM is called again on the current level of the hierarchy to potentially improve the objective function further, until a clustering for the finest level, i.e. the original graph, is obtained.

GVM is closely related to algorithms in the context of graph partitioning and has previously been used for modularity-based clustering without constraints [5, 22]. Neither approximation guarantees nor subexponential bounds on the running time are known, but experimentally it has been shown to outperform the corresponding greedy merge algorithm with respect to both quality and efficiency. For modularity, it can easily be shown that moving a vertex to a cluster it is not linked with is never the best choice, therefore it suffices to consider neighboring clusters. Together with the observation that the change in modularity can be determined in constant time for each move if some information

about the clustering is maintained, this yields a running time in $O(m)$ for each round in LM. This latter observation on running time also holds for all intracluster density and intercluster sparsity measures except for mixd, mixc and mixe, whose values are expensive to maintain.

**Ensuring Strict Improvements.** Another issue with a direct application of GVM to maximum-based measures is that iteratively traversing the whole vertex set is inefficient if only very few vertex moves potentially decrease the cut of the cluster with the currently worst value. Even worse, if this cluster is not unique, it is likely that the search is stuck in a local minimum, as vertex moves generally can only improve the value for one of these clusters, not for all of them simultaneously. If we try to prevent this by allowing vertex moves that are not strictly improving, we somehow have to ensure that the algorithm terminates after a finite number of operations. We do this in a similar way as proposed in [14] for GM by greedily optimizing the lexicographical order of the intercluster sparsity values of all the clusters. Let $L(\mathcal{C}) := \big(f(C_1), \ldots, f(C_k)\big), C_i \in \mathcal{C}$, be the sequence of these values with decreasing intercluster density, i.e. $(f(C_i) \geq f(C_{i+1})$ for $i \in \{1, \ldots, k-1\}$. Then a clustering $\mathcal{C}$ is *L-better* than $\mathcal{C}'$ if $L(\mathcal{C})$ is lexicographically smaller than $L(\mathcal{C}')$. We now determine for each vertex the set of clusterings that can be reached by moving it. If one of these clusterings is $L$-better than the current clustering, the move that results in the $L$-best sequence is performed. As we strictly improve the lexicographical order in each step, termination is guaranteed. This means, we greedily optimize the maximum value but are also allowed to improve the intercluster sparsity of clusters more locally, yielding better efficiency and the possibility to escape local minima.

**Determining the Best Move in $O(\deg(v))$ Time.** It holds that any two clusterings resulting from leaving vertex $v$ untouched or from moving $v$ to a different (or new) cluster can be $L$-compared in constant time (see App. A). Furthermore, it is immediate that moving a vertex to a cluster it is not linked to can never decrease the number of intercluster edges (nxe). This does not hold for gxd, however, it is not hard to see that GVM never has to consider non-neighboring clusters for gxd (see App. A). For all other intercluster density

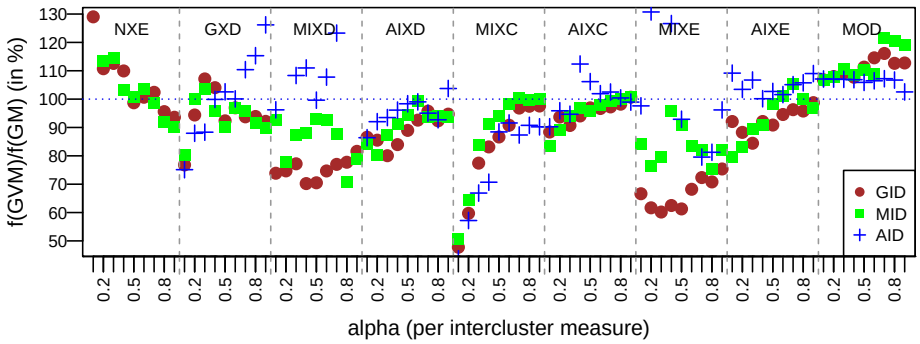



Fig. 1: Qualitative comparison of GVM and GM.

measures this does not hold as can be seen in the examples in Fig. 4 in App. A. As configurations like these are only expected in degenerate cases, the impact on efficiency is large on sparse graphs, and unconnected clusters are not desirable in the context of graph clustering, we chose to restrict the set of feasible moves to neighboring clusters. Together with the possibility to compare different moves in constant time, we get a time complexity of $O(m)$ for each round of the local move procedure for each of the combinations considered.

## 4 Experiments

**Qualitative Comparison of Greedy Merge and Greedy Vertex Moving.** Our first experiments address the question which flavor of greedy algorithm is better suited for DDC. As test instances, we used all graphs listed in Table 2 with less than 1000 vertices, these are real-world networks taken from the websites of Mark Newman [20] and Alex Arenas [3] and are part of the clustering testbed of the 10th DIMACS Implementation Challenge [1]. For all proposed combinations of measures, Figure 1 shows the ratio of the intercluster density obtained by using GVM and GM, averaged over all graphs. For modularity, this ratio is always greater than one, confirming that local moving yields better results, regardless of the choice and strength of the constraint. In combination with gid and mid, this similarly holds for all other objectives except for nxe, note that, in contrast to modularity, we aim to minimize these measures and therefore a value below one means that GVM attains better results. For nxe, the outcome depends on the value of $\alpha$ chosen. In combination with aid, the outcome is less clear, the results for nxe are out of bounds as the ratio for some configurations exceeds 300 percent. This can be explained by the observation that aid happily allows (and thereby encourages) unbalanced clusterings, as bad intracluster density values of large clusters can easily be compensated by a set of small and dense clusters, and GM is known to have a tendency to produce unbalanced partitions. As this most often leads to unintuitive clusterings, we deem aid less suitable in the context of graph clustering. Disregarding aid for these reasons, in a vast majority of configurations, GVM outperforms GM. For tackling DCC, we thus solely use GVM, putting aside the algorithm based on greedy merging.

**Effectiveness of Different Objective Functions.** The next question we pose is, if each of the intercluster density measures is effective in optimizing itself when used as inter in GVM. To answer this question, we conducted the following experiment on the set of graphs listed in Table 2. In the following, let $\text{GVM}_{i,\alpha,x}$ denote GVM incorporating the constraint $i(\mathcal{C}) \geq \alpha$ and the objective $x(\mathcal{C})$. For each setup of DDC, i.e. intracluster measure $i$, intercluster measure $x$ and $\alpha \in \{0.0, 0.1, \ldots, 1.0\}$, we ranked the clusterings obtained by $\text{GVM}_{i,\alpha,y}$ by their performance with respect to $x$, using all possible objectives $y$ for GVM. Figure 2 shows the distribution of these ranks over all configurations involving gid, grouped by $x$. The outcome of this experiment is less clear than what might be expected—none of the intercluster measures, not even modularity, scores the best quality with respect to itself in all configurations. Nonetheless, in general,

| graph | n | m | graph | n | m |
|---|---|---|---|---|---|
| karate(N) | 34 | 78 | netscience(N) | 1589 | 2742 |
| dolphins(N) | 62 | 159 | power(N) | 4941 | 6594 |
| lesmis(N) | 77 | 254 | hep-th(N) | 8361 | 15751 |
| polbooks(N) | 105 | 441 | PGPgiantcompo(A) | 10680 | 24316 |
| adjnoun(N) | 112 | 425 | astro-ph(N) | 16706 | 121251 |
| football(N) | 115 | 616 | cond-mat(N) | 16726 | 47594 |
| jazz(A) | 198 | 2742 | as-22july06(N) | 22963 | 48436 |
| celegansneural(N) | 297 | 2148 | cond-mat-2003(N) | 31163 | 120029 |
| celegans_metabolic(A) | 453 | 2039 | cond-mat-2005(N) | 40421 | 175693 |
| polblogs(N) | 1490 | 16718 | | | |

Table 2: List of the real world test instances ordered by increasing number of vertices. These are taken from the webpages of Arenas(A) [3] and Newman(N) [20] and are often used to compare clustering algorithms. All graphs are part of the clustering testbed of the 10th DIMACS Implementation Challenge [1].

except for nxe which is clearly dominated by gxd, each objective optimizes itself quite well. This also holds for mid, while for aid, the outcome is even less clear, as can be seen in Figures 5, 6 in App. B.

**Reference Algorithms.** For a more comprehensive assessment of GVM as a means to address DCC, we use the following reference algorithms:

- *Iterative Conductance Cutting (ICC)* [18]: This top-down algorithm iteratively splits the input graph into two subgraphs based on a cut with low conductance. The process stops when the conductance of the cut exceeds a given threshold, which we set to 0.4 in our experiments.
- *Markov-Clustering (MCL)* [10]: Emulating a random walk, the matrix of transition probabilities is alternately taken to to the power of $e$ and renormalized after taking each entry to the power of $r$, where $e$ and $r$ are input parameters. In our experiments, we set $r$ and $e$ to 2.
- *Geometric MST Clustering (GMC)* [6]: First, a spectral embedding of the graph in $d$-dimensional space is built. Then the algorithm constructs a Euclidean minimum spanning tree and successively deletes the heaviest edge. This defines a sequence of forests whose connected components induce a set of clusterings. Among these clusterings, the one with the best value according to some given objective function is chosen.
- *Multi-Level Modularity (MOD)* [22]: This is the GVM-algorithm based solely on modularity without using any constraint. This algorithm has been shown to perform very well in the context of Modularity optimization [22].

**Comparison Based on Intracluster Density Found by Reference Algorithms.** ICC, MCL and MOD do not incorporate constraints on the intracluster density of the resulting clustering. Nonetheless, it is still possible to evaluate them with respect to those variants of DCC, where $\alpha$ is set to the intracluster density found by these algorithms. In other words, given the same constraint a

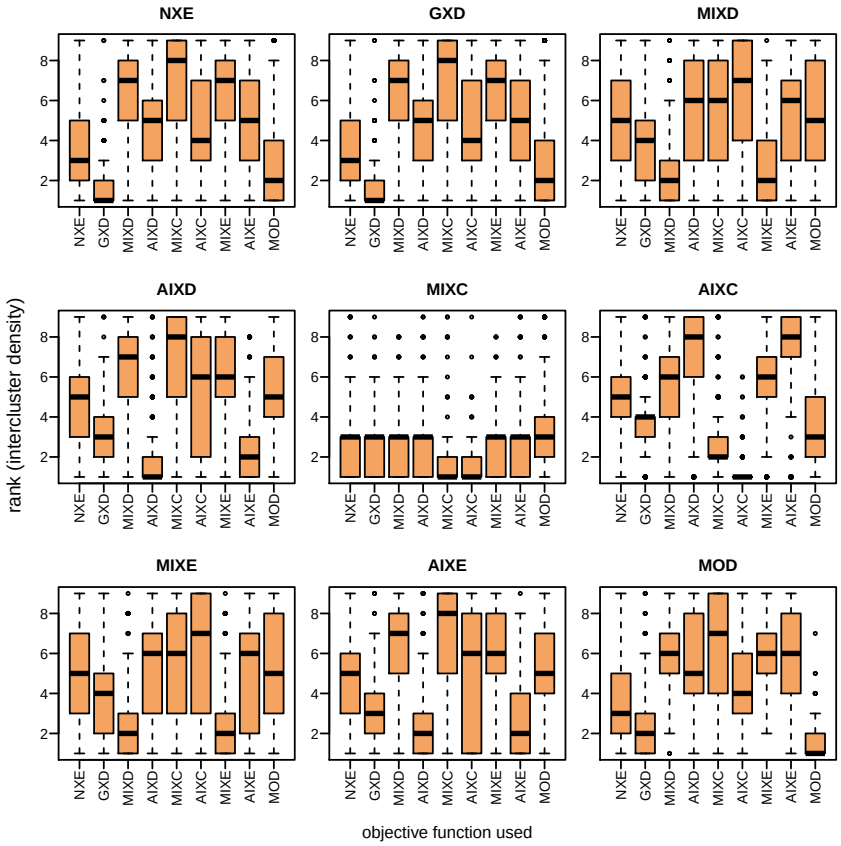



Fig. 2: Ranks for different intercluster density measures as objectives in the GVM-algorithm using gid as constraint, evaluated by the intercluster density of the resulting clustering.

reference algorithm $\mathcal{A}$ implicitly adheres to, how well does GVM compare to $\mathcal{A}$ wrt. DDC?

We first ran ICC, MCL and MOD on all test instances in Table 2 and recorded the intracluster density values $\alpha$ of the resulting clusterings. Then, for each reference algorithm $\mathcal{A}$, corresponding $\alpha$, $i$, and $x$, we compare the clustering obtained by $\text{GVM}_{i,\alpha,x}$ to the clustering of $\mathcal{A}$ with respect to $x$. For GMC the experiments slightly differ as GMC requires an objective function. We filled this degree of freedom by choosing $f(\mathcal{C}) = i(\mathcal{C}) - x(\mathcal{C})$ as the objective function for the experiments using $i$ as intracluster and $x$ as intercluster density measure. This seemed to be the fairest way of comparison and in almost all cases led to non-trivial clusterings.

Table 3 and Table 4 show the percentage of graphs where the greedy algorithm for $x$ compares favorably and the arithmetic mean of the ratio of $x$

obtained with GVM and with the reference algorithm. As we aim to minimize intercluster density, a value below 1 indicates that the greedy algorithm succeeds in beating the reference algorithm and vice versa. Compared to ICC and MCL, GVM clearly yields better results. The same holds for GMC, except if used in combination with aid, where GMC sometimes produces far better results. This can be explained by the fact that aid does not punish unbalancedness and GMC naturally leads to very unbalanced clusterings in most instances. The outcome of the comparison with the modularity-based algorithm is less clear. For aid, GVM performs better, which is not surprising as modularity strongly discourages unbalanced clusterings. For mid, GVM still beats MOD in the majority of configurations, while for gid, this only holds for slightly less than half of the configurations. Furthermore, it is worth mentioning that especially for aixd and aixe there are instances where modularity minimizes these functions far better than the respective greedy algorithms. Altogether, the comparison with ICC, MCL and GMC suggests that GVM effectively addresses DDC, while the comparison

| | gid | | | | mid | | | | aid | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ICC | MCL | MOD | GMC | ICC | MCL | MOD | GMC | ICC | MCL | MOD | GMC |
| nxe | 84 | 95 | 16 | 63 | 89 | 95 | 63 | 74 | 95 | 100 | 100 | 63 |
| gxd | 84 | 100 | 42 | 100 | 95 | 100 | 84 | 100 | 95 | 100 | 100 | 84 |
| aixd | 84 | 100 | 42 | 100 | 89 | 100 | 37 | 95 | 95 | 100 | 100 | 84 |
| aixc | 84 | 100 | 21 | 53 | 95 | 100 | 79 | 42 | 95 | 95 | 100 | 63 |
| aixe | 84 | 95 | 42 | 89 | 89 | 95 | 42 | 95 | 95 | 95 | 95 | 95 |
| mixd | 84 | 95 | 53 | 84 | 89 | 100 | 74 | 89 | 89 | 95 | 89 | 74 |
| mixc | 89 | 95 | 42 | 37 | 89 | 95 | 63 | 37 | 89 | 95 | 84 | 21 |
| mixe | 89 | 95 | 58 | 89 | 84 | 95 | 47 | 79 | 95 | 95 | 89 | 63 |

Table 3: Comparison of GVM and reference algorithms. Entries represents the percentage of graphs GVM compares favorably.

| | gid | | | | mid | | | | aid | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ICC | MCL | MOD | GMC | ICC | MCL | MOD | GMC | ICC | MCL | MOD | GMC |
| nxe | 0.67 | 0.52 | 1.17 | 1.26 | 0.42 | 0.08 | 0.97 | 0.88 | 0.03 | 0.06 | 0.05 | 8.07 |
| gxd | 0.64 | 0.50 | 1.07 | 0.11 | 0.40 | 0.09 | 0.89 | 0.10 | 0.07 | 0.10 | 0.13 | 0.76 |
| aixd | 0.47 | 0.32 | 5.30 | 0.25 | 0.34 | 0.06 | 5.08 | 0.23 | 0.18 | 0.12 | 0.22 | 0.61 |
| aixc | 0.57 | 0.29 | 2.17 | 0.28 | 0.39 | 0.05 | 0.81 | 0.27 | 0.41 | 0.27 | 0.37 | 7.87 |
| aixe | 0.49 | 0.39 | 5.55 | 0.31 | 0.36 | 0.14 | 5.22 | 0.31 | 0.19 | 0.13 | 0.24 | 1.45 |
| mixd | 0.45 | 0.34 | 0.96 | 0.41 | 0.39 | 0.07 | 1.27 | 0.30 | 0.21 | 0.18 | 0.32 | 3.17 |
| mixc | 0.69 | 0.58 | 1.15 | 0.34 | 0.47 | 0.15 | 1.09 | 0.30 | 0.44 | 0.39 | 0.46 | 1.60 |
| mixe | 0.48 | 0.26 | 1.25 | 0.57 | 0.39 | 0.14 | 1.28 | 0.63 | 0.13 | 0.16 | 0.28 | 3.02 |

Table 4: Comparison of GVM and reference algorithms. Entries represent the mean ratio of the respective intercluster measure $x$ obtained by GVM and reference algorithm.

with MOD shows that optimizing modularity is similarly effective in minimizing cut-based intercluster sparsity measures.

**Recovering Planted Partitions.** To compare the different objective functions qualitatively, we evaluated how well the corresponding GVM-algorithms are able to reconstruct planted partitions in random graphs. As a comparison, we also give the results obtained by MOD. Due to higher running times and large numbers of experiments, we omit a comparison with ICC, MCL and GMC.

**Random Graphs Generated.** We use an adapted Erdős-Rényi-model, where, starting from a given reference partition, the probability that vertices in the same set (in different sets) are connected equals $p_{in}$ ($p_{out}$). The number of vertices ($n$) and clusters ($k$) as well as the skewness of the distribution of cluster sizes ($\beta$) of the planted partition are input parameters. Setting $\beta$ to 1.0 corresponds to uniform cluster sizes, values below and above 1 cause this distribution to be skewed, for more details see [15]. As configurations, we fixed $n$ to 10000 and chose $p_{in}$ and $p_{out}$ such that the average number of intracluster (intercluster) edges a vertex is incident to equals 5 (3). To determine the reference partition, we used all combinations of $k \in \{10, 100, 300\}$ and $\beta \in \{0.3, 1.0, 2.0\}$. For each configuration, we generated 100 instances and always averaged obtained values.

**Distance Measures.** To compare the clusterings obtained with the different algorithms with the reference clustering, we use the following graph-based distance measures taken from [9]:

- *Graph-based Rand Index ($R_g$)*: Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be clusterings and $e_{11}$ ($e_{00}$) the number of edges which are intracluster (intercluster) wrt. both $\mathcal{C}_1$ and $\mathcal{C}_2$. Then, $R_g(\mathcal{C}_1, \mathcal{C}_2) = 1 - (e_{11} + e_{00})/m$.
- *Editing Set Difference (ESD)*: For a clustering $\mathcal{C}$, its editing set $F_{\mathcal{C}}$ is the set of edges requiring insertion or removal such that the clusters in $\mathcal{C}$ form disjoint cliques. Then, for clusterings $\mathcal{C}_1$ and $\mathcal{C}_2$, their editing set difference is defined as $ESD(\mathcal{C}_1, \mathcal{C}_2) = 1 - |F_{\mathcal{C}_1} \cap F_{\mathcal{C}_2}|/|F_{\mathcal{C}_1} \cup F_{\mathcal{C}_2}|$.

**Parameters and Evaluation.** As an exhaustive parameter search for all configurations would be far too expensive, we always set $\alpha$ to 75 percent of the expected global intracluster density $p_{\text{in}}$. We deemed taking the actual value of $p_{\text{in}}$ too strict, as, especially for mid, even the reference clustering of the generator most likely does not meet this constraint. The previous experiments indicate that there are configurations where particular objective functions used in GVM do not score the best results with respect to themselves. As our goal is to compare good clusterings with respect to different combinations of $i$ and $x$, independent of artifacts of GVM, we chose the following approach: For a combination $i$, $\alpha$, $x$, we evaluated the clustering that, among all results obtained with GVM using $i \geq \alpha$ as constraint, is best with respect to $x$ (as opposed to simply evaluating $\text{GVM}_{i,\alpha,x}$). Furthermore, preliminary experiments confirmed that constraining aid leads to very unintuitive and unbalanced clusterings, which is mirrored by the fact that the corresponding versions of DCC are far less effective in finding the hidden clustering. We hence excluded aid in the discussion of the results.

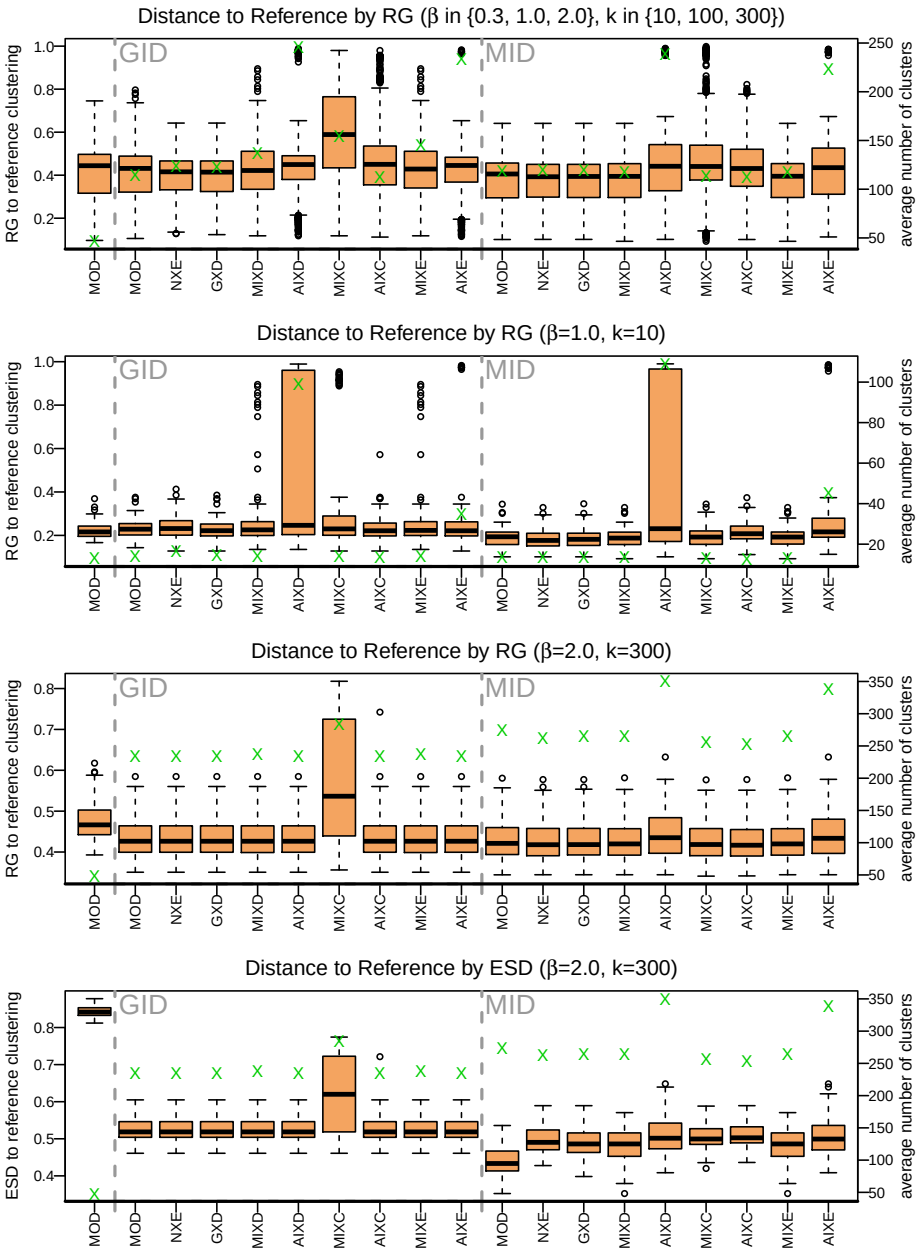



Fig. 3: Distance to reference clustering (boxplots, left-hand $y$-axis) and number of clusters discovered in planted partition graphs (green $\times$-marks, right-hand $y$-axis), different configurations

**Results on Planted Partition Graphs.** Figure 3 shows the results for selected configurations, the results for the whole set of experiments can be found in App. C. In the first plot it can be seen that, in general, the clusterings that are ranked best with respect to mod, nxe and gxd are most similar to the reference.

*Constraining modularity by* mid *improves its results.* This especially holds for the experiments with high skewness ($\beta = 2$) and $k = 300$. In these experiments, modularity finds far less clusters than expected, partially due to its known resolution limit [13], which can be circumvented by steering the coarseness of the clustering by constraining the intracluster density. Another interesting fact is that ESD punishes these coarse clusterings far more than $R_g$.

*Fine reference clusterings disbalance maximum objectives.* Compared to the above, especially mixc in combination with gid yields worse similarity values. This, and the slightly increased cluster count can be explained by a tendency of mixc to favor unbalanced clusterings if the expected number of clusters is high ($k = 300$), which also explains why this effect does not happen in combination with mid that does not allow very unbalanced clusterings. To a smaller extent, the same observation also holds for the other maximum measures, as can be seen for $k = 300$ and $\beta = 1.0$.

aixe *and especially* aixd *identify many clusters.* Another striking observation is that the average number of clusters in clusterings found by aixd and aixe, indicated by the green $\times$-marks, is much higher than the average number of clusters in the reference. This especially stems from the experiments with few clusters. In the configuration with $\beta = 1$ and $k = 10$, it can also be seen that these measures differ the more, the coarser the expected clustering gets. This is not unexpected, as the denominator of aixd grows more slowly with the number of vertices in the cluster than the denominator of aixe, meaning that aixd is less eager to produce very large clusters. Additionally, in [14] it was proven that with the exception of aixd, all intercluster measures considered here can always be ameliorated by merging two existing clusters (unboundedness), which is also a hint that aixd is less likely to produce coarse clusterings than the other measures.

**Implementation and Running Times.** The algorithms ICC, MCL, GMC and GM are implemented in Java 1.6.0_22 using the graph library yFiles [23]. GVM (also incorporating MOD as a special case) is implemented in C++ using version 1_42 of the Boost Graph Library [2] and compiled with gcc 4.5.2 with optimization level 4. The focus of this evaluation is on the quality of the resulting clusterings, not on running times. However, to get a rough impression about the latter, clustering cond-mat-2005 on a 2.1 GHz AMD Opteron processor takes about 6 hours with ICC, 1 hour and 50 minutes with MCL, 5 minutes with GMC and 3 to 15 seconds with GVM, depending on the parameter setting. With our prototype implementation (not including the improvements proposed in [14]) of GM, clustering the much smaller celegans_metabolic takes over 2 minutes.

## 5 Conclusion

This work is an experimental evaluation of algorithms for the optimization problem Density-Constrained Clustering (DDC). We first evaluated two greedy heuristics, vertex moving and cluster merging, against each other and against algorithms from the literature. Vertex moving proved reliably superior to cluster merging and, in many cases, beats the results of the reference algorithms. Our results also show that a well-known modularity-based algorithm implicitly addresses DCC quite well, revealing similarities between cut-based intercluster sparsity measures and modularity. In the second part, we addressed the question whether different combinations of intracluster density and intercluster sparsity measures are suitable to guide algorithms in recovering planted partitions in random graphs. The results suggest that minimizing the average intercluster expansion or density of the clusters overestimates the number of clusters if the expected clustering is coarse, while the maximum intercluster measures lead to unbalanced clusters if the expected clustering is fine and the constraint on the intracluster density does not force the clustering to be balanced. Additionally, it can be seen that the known resolution limit for modularity can be circumvented if the coarseness of the clustering is controlled by an additional constraint on the intracluster density of the clustering.

## References


1. 10th DIMACS implementation challenge, http://www.cc.gatech.edu/dimacs10/
2. Boost C++ Libraries, http://www.boost.org/
3. Arenas, A.: Network data sets, http://deim.urv.cat/~aarenas/data/welcome.htm
4. Berkhin, P.: A Survey of Clustering Data Mining Techniques. In: Kogan, J., Nicholas, C., Teboulle, M. (eds.) Grouping Multidimensional Data: Recent Advances in Clustering, pp. 25–71. Springer (2006), http://www.springerlink.com/content/x321256p66512121/
5. Blondel, V., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10) (2008), http://dx.doi.org/10.1088/1742-5468/2008/10/P10008
6. Brandes, U., Gaertler, M., Wagner, D.: Experiments on Graph Clustering Algorithms. In: Proceedings of the 11th Annual European Symposium on Algorithms (ESA'03). Lecture Notes in Computer Science, vol. 2832, pp. 568–579. Springer (2003), http://www.springerlink.com/openurl.asp?genre=article&issn=0302-9743&volume=2832&spage=568
7. Brandes, U., Gaertler, M., Wagner, D.: Engineering Graph Clustering: Models and Experimental Evaluation. ACM Journal of Experimental Algorithmics 12(1.1), 1–26 (2007), http://portal.acm.org/citation.cfm?id=1227161.1227162
8. Clauset, A., Newman, M.E.J., Moore, C.: Finding community structure in very large networks. Physical Review E 70(066111) (2004), http://link.aps.org/abstract/PRE/v70/e066111
9. Delling, D., Gaertler, M., Görke, R., Wagner, D.: Engineering Comparators for Graph Clusterings. In: Proceedings of the 4th International Conference on Algorithmic Aspects in Information and Management (AAIM'08). Lecture Notes in Computer Science, vol. 5034, pp. 131–142. Springer (June 2008)

10. van Dongen, S.M.: Graph Clustering by Flow Simulation. Ph.D. thesis, University of Utrecht (2000), http://micans.org/mcl/lit/
11. Flake, G.W., Tarjan, R.E., Tsioutsiouliklis, K.: Graph Clustering and Minimum Cut Trees. Internet Mathematics 1(4), 385–408 (2004), http://www.internetmathematics.org/volumes/1.htm
12. Fortunato, S.: Community detection in graphs. Physics Reports 486(3–5), 75–174 (2010), http://www.sciencedirect.com/science/journal/03701573
13. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Science of the United States of America 104(1), 36–41 (2007), http://www.pnas.org/content/104/1/36.full.pdf
14. Görke, R., Schumm, A., Wagner, D.: Density-Constrained Graph Clustering. In: Dehne, F., Iacono, J., Sack, J.R. (eds.) Algorithms and Data Structures, 12th International Symposium (WADS'11). Lecture Notes in Computer Science, vol. 6844. Springer (August 2011)
15. Görke, R., Staudt, C.: A Generator for Dynamic Clustered Random Graphs. Tech. rep., ITI Wagner, Faculty of Informatics, Universität Karlsruhe (TH) (2009), http://i11www.iti.uni-karlsruhe.de/projects/spp1307/dyngen, informatik, Uni Karlsruhe, TR 2009-7
16. Hoory, S., Linial, N., Wigderson, A.: Expander graphs and their applications. Bulletin of the American Mathematical Society 43, 439–561 (2006)
17. Jain, A.K., Dubes, R.C.: Algorithms for Clustering Data. Prentice Hall (1988)
18. Kannan, R., Vempala, S., Vetta, A.: On Clusterings - Good, Bad and Spectral. In: Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science (FOCS'00). pp. 367–378 (2000)
19. Lancichinetti, A., Fortunato, S.: Community detection algorithms: A comparative analysis. Physical Review E 80(5) (November 2009), http://link.aps.org/doi/10.1103/PhysRevE.80.056117
20. Newman, M.: Network data, http://www-personal.umich.edu/~mejn/netdata/
21. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review E 69(026113), 1–16 (2004), http://link.aps.org/abstract/PRE/v69/e026113
22. Rotta, R., Noack, A.: Multilevel local search algorithms for modularity clustering. ACM Journal of Experimental Algorithmics 16, 2.3:2.1–2.3:2. (July 2011), http://doi.acm.org/10.1145/1963190.1970376
23. yWorks GmbH: yFiles for Java (2008), http://www.yworks.com/en/products_yfiles_about.html

# A Additional Examples and Explanations

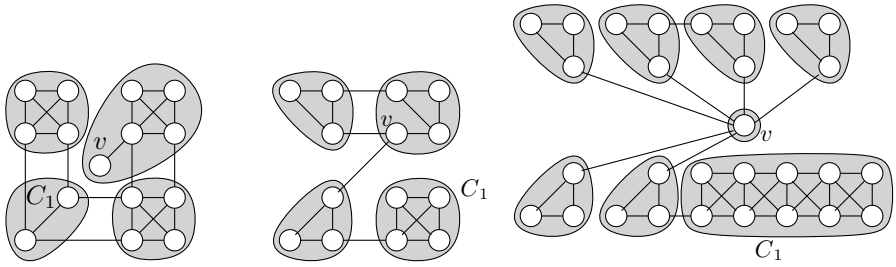
**Maximum Functions: Clusterings resulting from vertex moves can be $L$-compared in constant time.** *For three distinct clusters $C$, $A$ and $B$ in $\mathcal{C}$ and $v \in C$ it holds that:*

- $\mathcal{C}_{v \to A}$ *is $L$-better than* $\mathcal{C} \Leftrightarrow \big\{C \setminus \{v\}, A \cup \{v\}\big\}$ *is $L$-better than* $\{C, A\}$
- $\mathcal{C}_{v \to A}$ *is $L$-better than* $\mathcal{C}_{v \to B} \Leftrightarrow \big\{A \cup \{v\}, B\big\}$ *is $L$-better than* $\big\{B \cup \{v\}, A\big\}$

If the volume, size and number of out-going edges of the clusters $A$, $B$ and $C$ are maintained by the algorithm, the density/conductance/expansion of $C, A, B, C \setminus \{v\}, A \cup \{v\}$ and $B \cup \{v\}$ can be determined in constant time. Hence, the conditions on the right-hand side can be evaluated in constant time, which can be used to determine the best move for a vertex efficiently.

**Connectedness of gxd.** The following equation shows that GVM never has to consider non-neighboring clusters for gxd, as isolating the respective vertex is always more beneficial. Let $v \in V$, $A := C(v) \setminus \{v\}$ and $B \in \mathcal{C}$ such that $m_{\{v\},B} = 0$, then:

$$\begin{aligned}\mathsf{gxd}(\mathcal{C}_{v \to \{\}}) &= \frac{\sum_{C_i, C_j, j>i} m_{C_i, C_j} + m_{\{v\},A}}{\sum_{C_i, C_j, j>i} |C_i||C_j| + |A|} \\ &< \frac{\sum_{C_i, C_j, j>i} m_{C_i, C_j} + m_{\{v\},A} - \overbrace{m_{\{v\},B}}^{=0}}{\sum_{C_i, C_j, j>i} |C_i||C_j| + |A| - \underbrace{|B|}_{>0}} = \mathsf{gxd}(\mathcal{C}_{v \to B})\end{aligned}$$




(a) mixd, mixe (b) mixc (c) aixc, aixe, aixd

Fig. 4: Examples illustrating that most measures considered do not enforce connected moves. Given the clusterings indicated by the gray areas, among all moves involving $v$, moving $v$ to cluster $C_1$ yields the largest decrease in the objective function.

## B Effectiveness of Different Objective Functions: Additional Plots

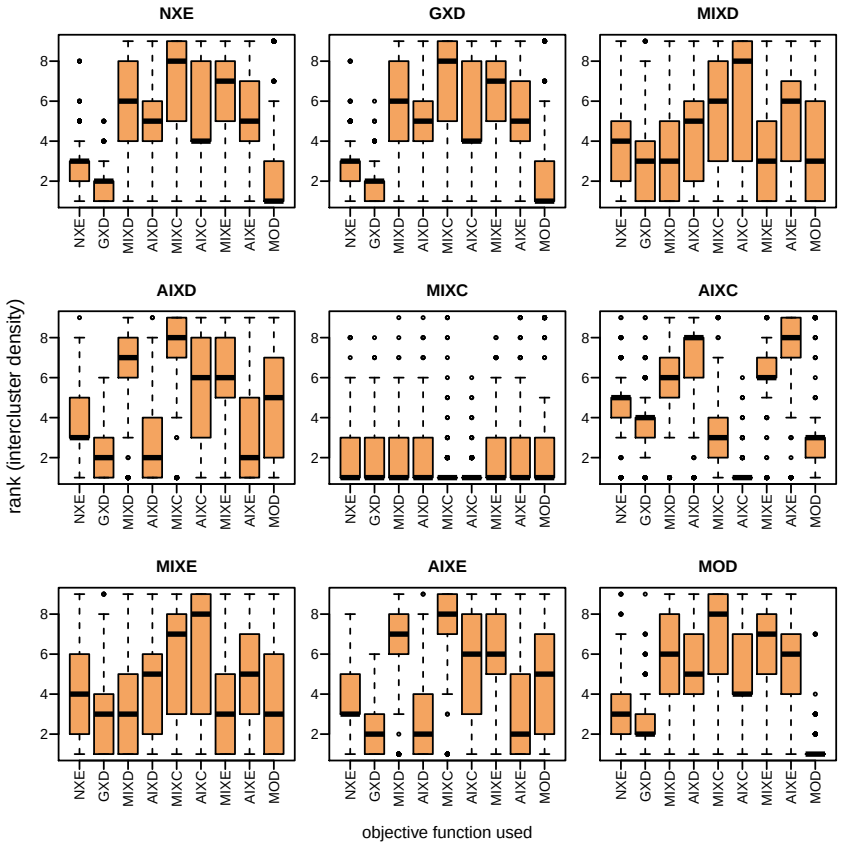



Fig. 5: Ranks for different intercluster density measures as objectives in the GVM-algorithm using mid as constraint, evaluated by the intercluster density of the resulting clustering.

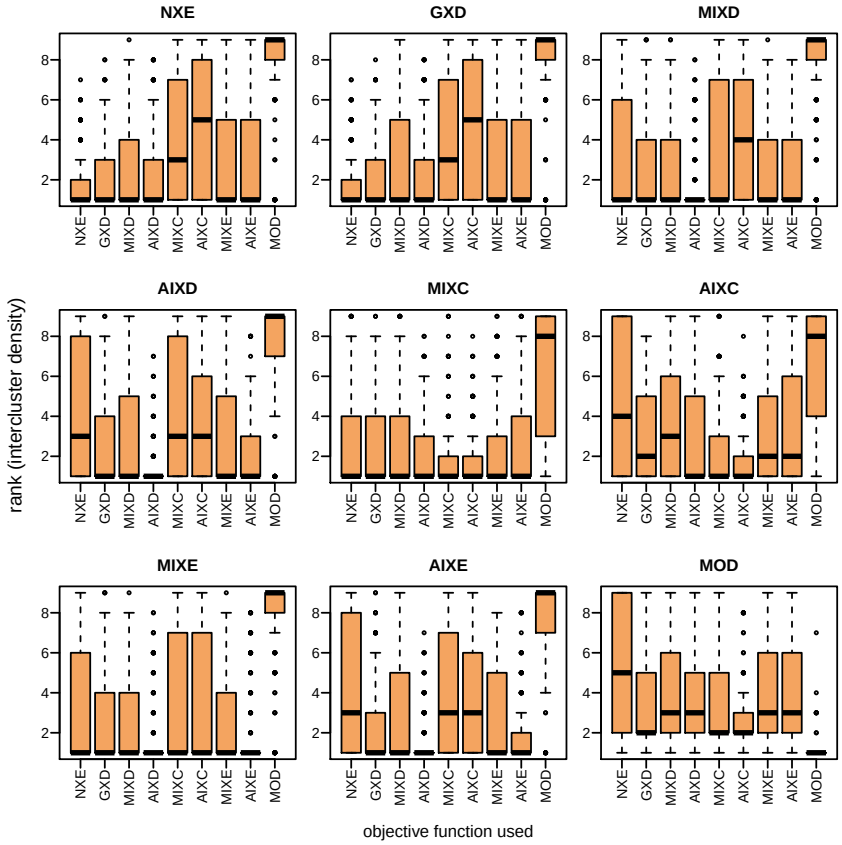
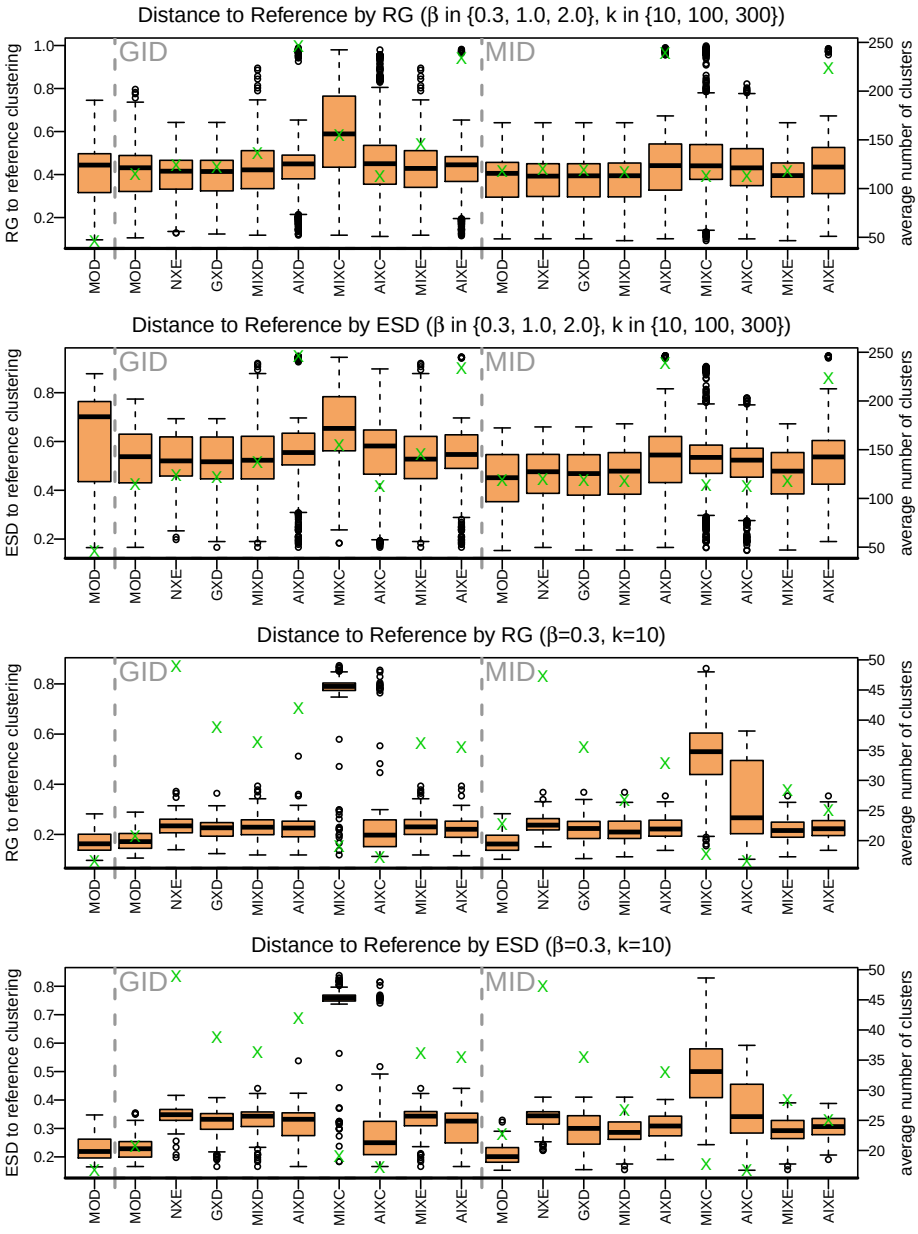
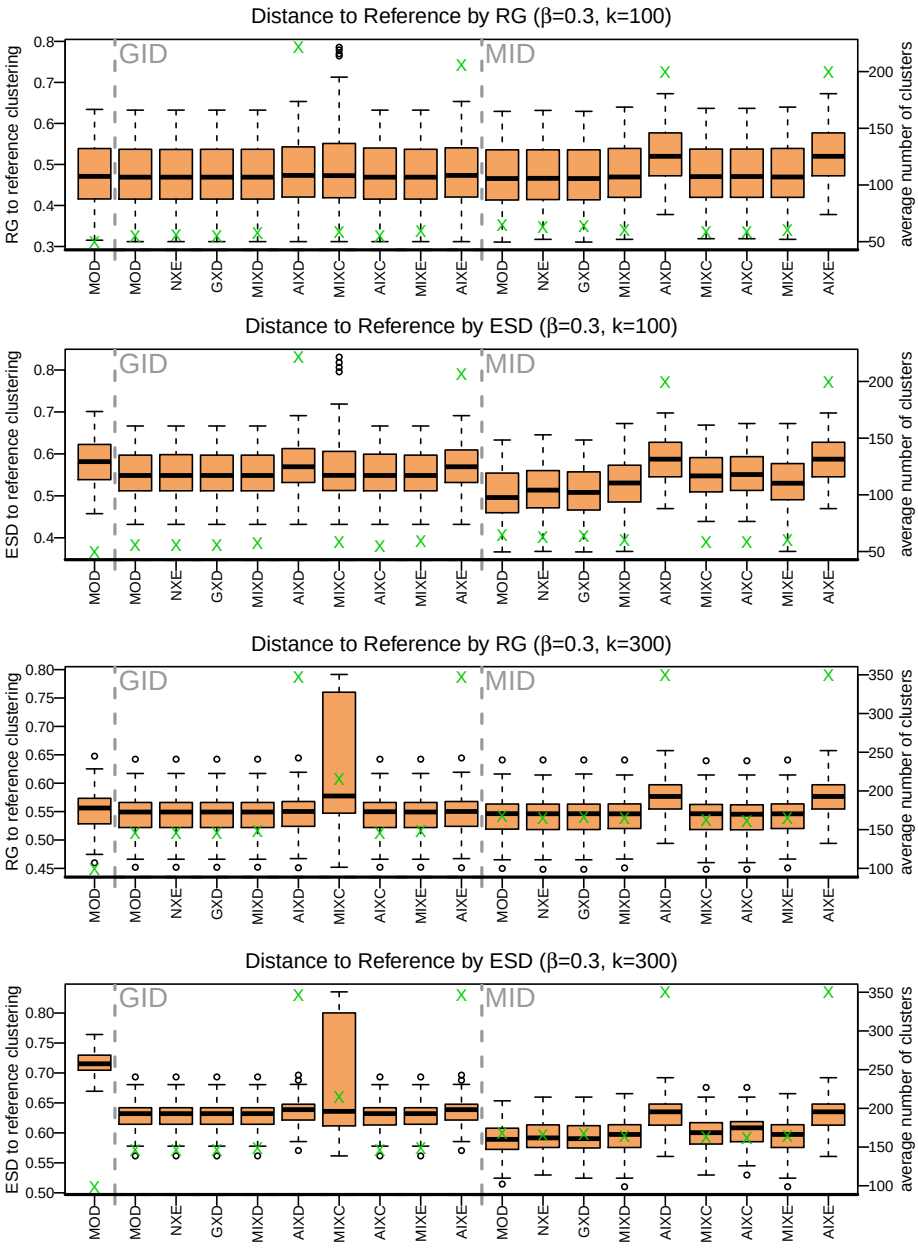
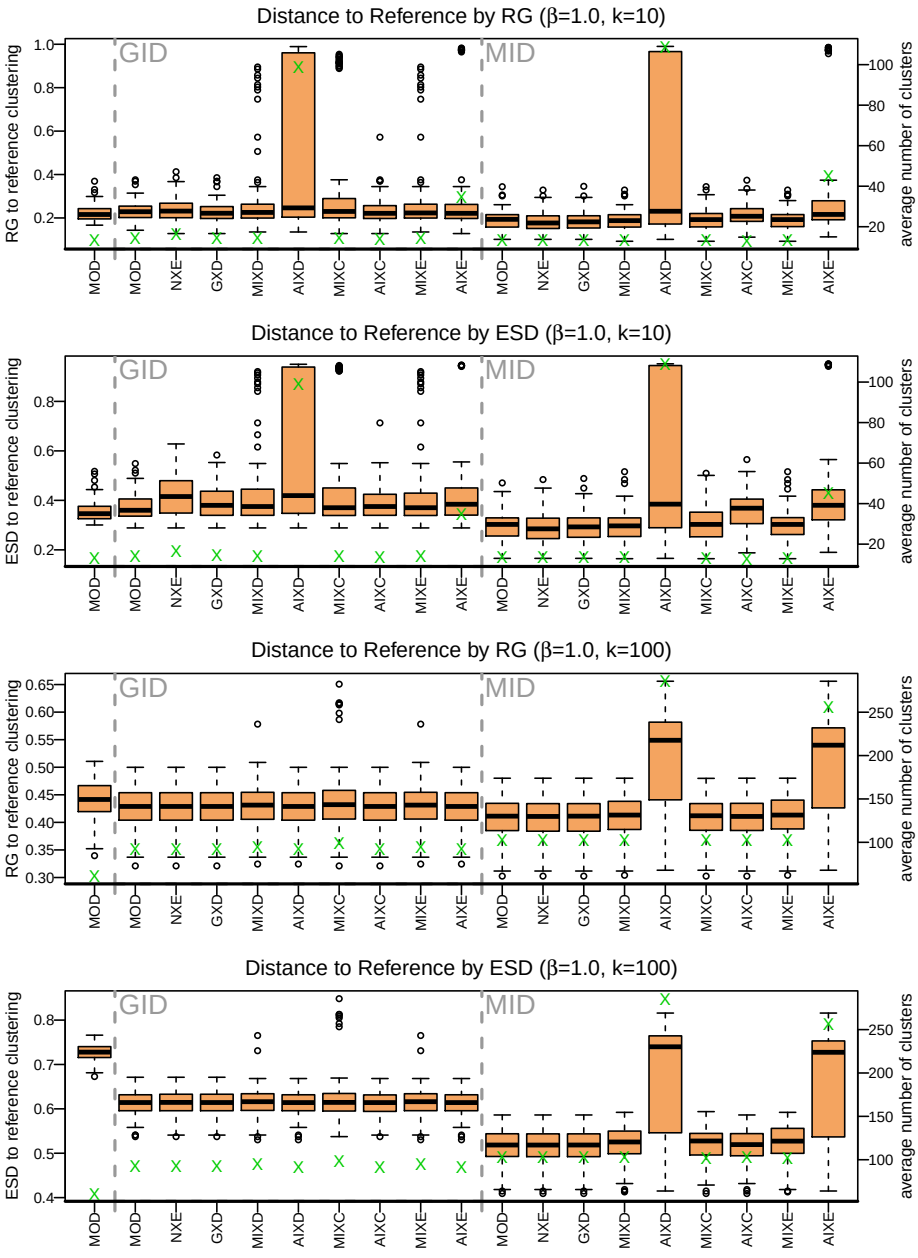


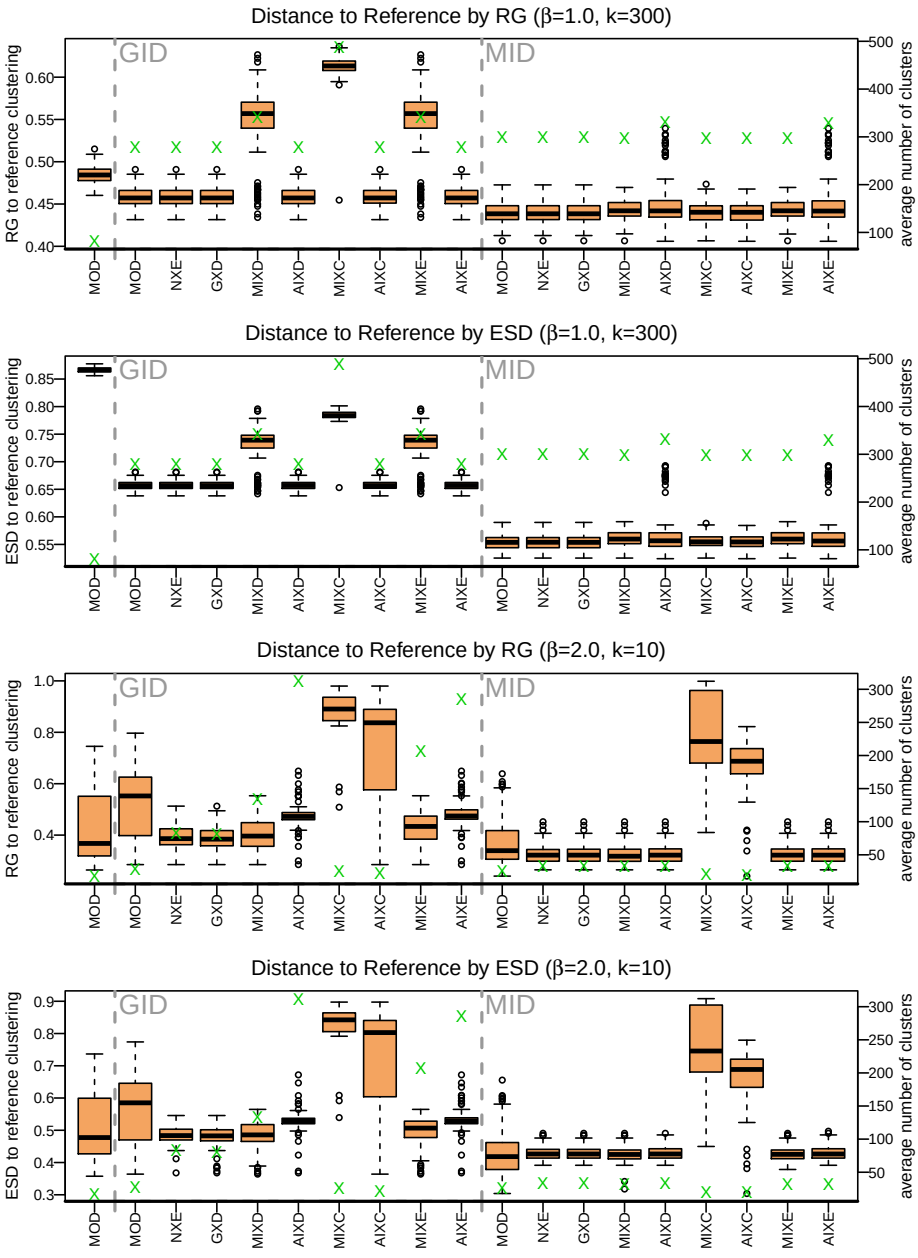

Fig. 6: Ranks for different intercluster density measures as objectives in the GVM-algorithm using aid as constraint, evaluated by the intercluster density of the resulting clustering.

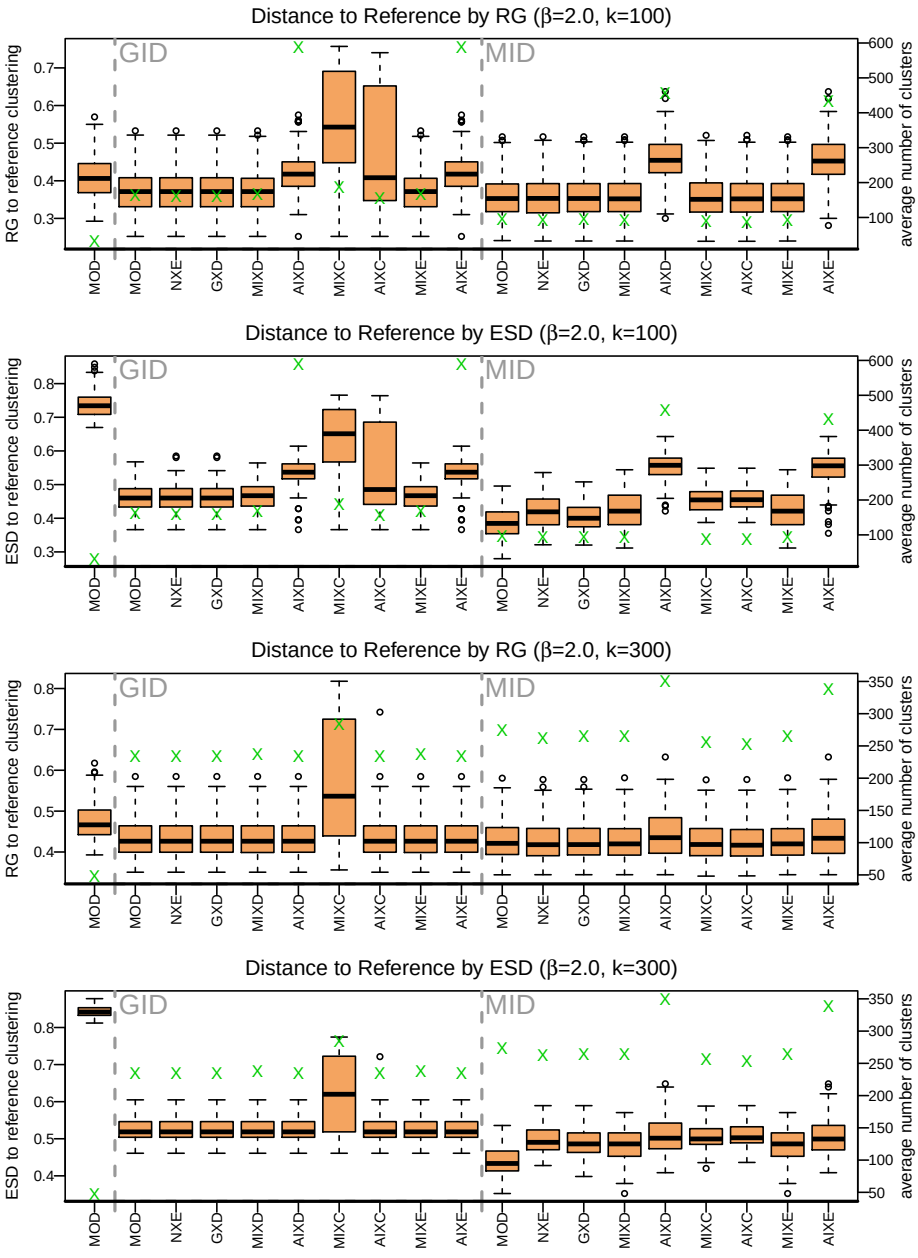
# C Complete Experiments with Planted Partition Graphs




Distance to Reference by RG (β=0.3, k=100)
RG to reference clustering
average number of clusters
GID
MID
Distance to Reference by ESD (β=0.3, k=100)
ESD to reference clustering
average number of clusters
GID
MID
Distance to Reference by RG (β=0.3, k=300)
RG to reference clustering
average number of clusters
GID
MID
Distance to Reference by ESD (β=0.3, k=300)
ESD to reference clustering
average number of clusters
GID
MID
MOD
NXE
GXD
MIXD
AIXD
MIXC
AIXC
MIXE
AIXE


Distance to Reference by RG (β=1.0, k=10)
RG to reference clustering
average number of clusters
GID
MID
MOD
NXE
GXD
MIXD
AIXD
MIXC
AIXC
MIXE
AIXE
Distance to Reference by ESD (β=1.0, k=10)
ESD to reference clustering
Distance to Reference by RG (β=1.0, k=100)
Distance to Reference by ESD (β=1.0, k=100)


Distance to Reference by RG (β=1.0, k=300)
RG to reference clustering
average number of clusters
GID
MID
Distance to Reference by ESD (β=1.0, k=300)
ESD to reference clustering
average number of clusters
GID
MID
Distance to Reference by RG (β=2.0, k=10)
RG to reference clustering
average number of clusters
GID
MID
Distance to Reference by ESD (β=2.0, k=10)
ESD to reference clustering
average number of clusters
GID
MID
MOD
NXE
GXD
MIXD
AIXD
MIXC
AIXC
MIXE
AIXE


Distance to Reference by RG (β=2.0, k=100)
RG to reference clustering
average number of clusters
GID
MID
Distance to Reference by ESD (β=2.0, k=100)
ESD to reference clustering
average number of clusters
GID
MID
Distance to Reference by RG (β=2.0, k=300)
RG to reference clustering
average number of clusters
GID
MID
Distance to Reference by ESD (β=2.0, k=300)
ESD to reference clustering
average number of clusters
GID
MID
MOD
NXE
GXD
MIXD
AIXD
MIXC
AIXC
MIXE
AIXE