

Correlation Clustering: from Local to Global Constraints

(Discussion Paper)

Domenico Mandaglio¹, Andrea Tagarelli¹ and Francesco Gullo²

¹*DIMES Dept., University of Calabria, 87036 Rende (CS), Italy*

²*UniCredit, Rome, Italy*

Abstract

Given a set of data objects, consider that object pairs are assigned two weights expressing the advantage of putting those objects in the same cluster or in separate clusters, respectively. Correlation clustering partitions the input object set so as to minimize the sum of the intra-cluster negative-type weights plus the sum of the inter-cluster positive-type weights. Existing approximation algorithms provide quality guarantees if the weights are bounded in some way. Regardless of the type, the weight bounds that have been so far studied are *local bounds*, i.e., constraints that are required to hold for every object pair in isolation. In this paper, we discuss *global weight bounds* in correlation clustering, and in particular, we derive bounds on edge weights' aggregate functions that are sufficient to lead to proved quality guarantees. Our formulation extends the range of applicability of the most prominent existing correlation-clustering algorithms thus providing benefits, both theoretical and practical. Also, we showcase our results in a real-world scenario of feature selection for fair clustering.

Keywords

min-disagreement correlation clustering, probability constraint, fair clustering

1. Introduction

Correlation clustering [1] is an important clustering formulation that has received considerable attention from theoreticians and practitioners, and found application in several contexts [2].

Given a set of objects and nonnegative real weights expressing “positive” and “negative” feeling of clustering any two objects together, *min-disagreement correlation clustering* (MIN-CC) partitions the input object set so as to minimize the sum of the intra-cluster negative-type weights plus the sum of the inter-cluster positive-type weights. Min-disagreement correlation clustering is **APX**-hard, but efficient constant-factor approximation algorithms exist if the weights are bounded in some way. The weight bounds tackled so far in the literature are said *local*, as they are required to hold *for every single object pair*.

In this paper, we discuss the main theoretical and experimental results from our study in [3], where we introduced the problem of min-disagreement correlation clustering with *global weight bounds*, i.e., constraints to be satisfied by the input weights altogether. Our main contribution is a sufficient condition that establishes when any algorithm achieving a certain


SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19-22, 2022, Tirrenia (PI), Italy

✉ d.mandaglio@dimes.unical.it (D. Mandaglio); andrea.tagarelli@unical.it (A. Tagarelli);

francesco.gullo@unicredit.eu (F. Gullo)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Algorithm 1 Pivot [4]

Input: Graph $G = (V, E)$; nonnegative weights $w_e^+, w_e^-, \forall e \in E$

Output: Clustering \mathcal{C} of V

- 1: $\mathcal{C} \leftarrow \emptyset, V' \leftarrow V$
 - 2: **while** $V' \neq \emptyset$ **do**
 - 3: pick a pivot vertex $u \in V'$ uniformly at random
 - 4: add $\mathcal{C}_u = \{u\} \cup \{v \in V' \mid (u, v) \in E, w_{uv}^+ > w_{uv}^-\}$ to \mathcal{C} and remove \mathcal{C}_u from V'
-

approximation under the probability constraint keeps the same guarantee on an input that violates the constraint. This extends the range of applicability of the most prominent existing correlation-clustering algorithms, including the popular Pivot, thus providing both theoretical and practical benefits. Experiments have shown the usefulness of our approach, in terms of both worthiness of employing existing efficient algorithms, and guidance on the definition of weights from feature vectors in a task of *fair clustering*.

2. Correlation Clustering with local weight bounds

The input of correlation clustering is a set V of objects, and two nonnegative, real-valued weights w_{uv}^+, w_{uv}^- for every (unordered) object pair $u, v \in V$. Any “positive” w_{uv}^+ (resp. “negative” w_{uv}^-) weight expresses the benefit of clustering u and v together (resp. separately). This input can equivalently be represented as a graph G with vertex set V and edge weights w_{uv}^+, w_{uv}^- , for all $u, v \in V$, and with edge (u, v) being drawn only if at least one among w_{uv}^+ and w_{uv}^- is nonzero.

In this work we tackle the problem of *min-disagreement correlation clustering*:

Problem 1 (MIN-CC [4]). *Given an undirected graph $G = (V, E)$, with vertex set V and edge set $E \subseteq V \times V$, and nonnegative weights $w_e^+, w_e^- \in \mathbb{R}_0^+$ for all edges $e \in E$, find a clustering (i.e., an injective function expressing cluster-membership) $\mathcal{C} : V \rightarrow \mathbb{N}^+$ that minimizes*

$$\sum_{(u,v) \in E, \mathcal{C}(u)=\mathcal{C}(v)} w_{uv}^- + \sum_{(u,v) \in E, \mathcal{C}(u) \neq \mathcal{C}(v)} w_{uv}^+. \quad (1)$$

For the sake of presentation, we assume $w_e^+ = w_e^- = 0$, for all $e \notin E$, and non-trivial MIN-CC instances, i.e., $w_e^+ \neq w_e^-$, for some $e \in E$.

MIN-CC is **NP**-hard [1, 5] yet difficult to approximate, being it **APX**-hard even for complete graphs and edge weights $(w_e^+, w_e^-) \in \{(0, 1), (1, 0)\}, \forall e \in E$ [6]. For general (i.e., not necessarily complete) graphs and unconstrained weights, the best known approximation factor is $\mathcal{O}(\log |V|)$ [6, 7]. This factor improves if restrictions on edge weights are imposed. The one that has received remarkable attention is the **probability constraint (PC)**: A MIN-CC instance is said to satisfy the *probability constraint* if $w_{uv}^+ + w_{uv}^- = 1$, for all pairs $u, v \in V$.

A MIN-CC instance obeying the PC necessarily corresponds to a complete graph (otherwise, any missing edge would violate the PC). Under the PC, MIN-CC admits constant-factor guarantees. The best known approximation factor is 4, achievable – as shown in [8] – by Charikar *et al.*’s algorithm [6]. That algorithm is based on rounding the solution to a large linear program (with a number $\Omega(|V|^3)$ of constraints), thus being feasible only on small graphs.

Algorithm 2 GlobalCC

Input: Graph $G = (V, E)$; nonnegative weights $w_e^+, w_e^-, \forall e \in E$, satisfying Theorem 1; algorithm A achieving α -approximation guarantee for MIN-PC-CC

Output: Clustering \mathcal{C} of V

- 1: choose M, γ s.t. $\frac{M}{\gamma} \in [\Delta_{max}, avg^+ + avg^-]$ {Theorem 1}
 - 2: compute $\sigma_{uv} = \gamma(w_{uv}^+ + w_{uv}^-) - M, \forall u, v \in V$
 - 3: compute $\tau_{uv}^\pm = \frac{1}{M} (\gamma w_{uv}^\pm - \frac{\sigma_{uv}}{2}), \forall u, v \in V$ (using M, γ defined in Step 1)
 - 4: $\mathcal{C} \leftarrow$ run A on MIN-PC-CC instance $\langle G' = (V, V \times V), \{\tau_e^+, \tau_e^-\}_{e \in V \times V} \rangle$
-

Here, we are particularly interested in the Pivot algorithm [4], due to its theoretical properties – it achieves a factor-5 expected guarantee for MIN-CC under the PC – and practical benefits – it takes $\mathcal{O}(|E|)$ time, and is easy-to-implement. Pivot simply picks a random vertex u , builds a cluster as composed of u and all the vertices v such that an edge with $w_{uv}^+ > w_{uv}^-$ exists, and removes that cluster. The process is repeated until the graph has become empty (Algorithm 1).

3. Correlation Clustering with global weight bounds

The weight bounds that have been so far studied are *local bounds*, i.e., constraints that are required to hold *for every object pair in isolation*. In this work, we are the first to consider *global weight bounds* in min-disagreement correlation clustering. We derive bounds on edge weights' aggregate functions that are sufficient to lead to proved quality guarantees. More specifically, for a MIN-CC instance $\langle G = (V, E), \{w_e^+, w_e^-\}_{e \in E} \rangle$ we define:

$$avg^+ = \binom{|V|}{2}^{-1} \sum_{e \in E} w_e^+, \quad avg^- = \binom{|V|}{2}^{-1} \sum_{e \in E} w_e^-, \quad \Delta_{max} = \max_{e \in E} |w_e^+ - w_e^-|$$

The main theoretical result of this work is described by the following theorem.

Theorem 1. *If the condition $avg^+ + avg^- \geq \Delta_{max}$ holds for a MIN-CC instance I , then it is possible to construct a MIN-CC instance I' (in linear time and space) such that (i) the probability constraint holds on I' , and (ii) an α -approximate clustering on I' (i.e., a clustering whose objective-function value is no more than α times I' 's optimum) is an α -approximate clustering on I too.*

Let MIN-PC-CC denote the version of MIN-CC operating on instances that satisfy the PC. According to Theorem 1, if $avg^+ + avg^- \geq \Delta_{max}$ holds for a MIN-CC instance, then any α -approximation algorithm A for MIN-PC-CC can be employed – *as a black box* – to get an α -approximate solution to that MIN-CC instance. The algorithm for doing so is simple: given a MIN-CC instance I , get a MIN-PC-CC instance via strict approximation-preserving (SAP) reduction, and run the black-box algorithm A on it (Algorithm 2). Being the reduction SAP, Algorithm 2 on input $\langle I, A \rangle$ achieves factor- α guarantee on I .

A noteworthy consequence of this result is that, if a MIN-CC instance I satisfies our condition, then the Pivot algorithm can be used to get (in linear time and space) a clustering achieving a 5-approximation guarantee on I .¹ This corresponds to extending the range of validity of

¹A probability-constraint-compliant MIN-CC instance I' is derivable from I in linear time and space (cf. Th. 1 (i)). Pivot on I' yields a 5-approximate clustering [4]. A 5-approximate clustering on I' is a 5-approximate clustering on I (cf. Th. 1 (ii)).

Pivot’s guarantee beyond the probability constraint: our global-weight-bounds condition now suffices for the 5-approximation to hold. A key advantage is that our condition is more likely to be satisfied than the probability constraint; for instance, it may happen that a bunch of edges are missing in the graph (thus violating the probability constraint), but, if our condition holds, one can get a 5-approximate clustering with Pivot. We point out that our result is general and holds for *any* MIN-CC algorithm achieving approximation guarantees under the probability constraint. However, the contextualization to the Pivot algorithm is relevant and worth to be exploited, since Pivot achieves the best tradeoff between quality guarantees and efficiency.

4. Analysis of the global-weight-bounds condition

Our result can be exploited to quickly yet easily recognize whether employing probability-constraint-aware approximation algorithms is a worth choice even if the probability constraint is not met. As an example, consider a graph that violates the probability constraint. So far, that graph would have likely been handled with linear-programming (LP) algorithms [6, 7], as they achieve (factor- $\mathcal{O}(\log |V|)$) approximation guarantees on general graphs/weights (whereas algorithms like Pivot are just heuristics if the probability constraint does not hold). Instead, our condition can be used as an indicator of whether Pivot can still achieve guarantees even if the probability constraint is violated, thus being preferred over the less efficient LP algorithms. Our experiments confirmed this theoretical finding, i.e. a better fulfilment of our condition corresponds to better performance of Pivot with respect to the LP algorithms, and vice versa.

Settings. We selected four real-world graphs, namely Karate, Dolphins, Adjnoun, and Football.² Note that the small size of such graphs is not an issue because this evaluation stage involves, among others, LP correlation-clustering algorithms, whose $\Omega(|V|^3)$ time complexity makes them unaffordable for graphs larger than that. We augmented these graphs with artificially-generated edge weights, to test different levels of fulfilment of our global-weight-bounds condition stated in Theorem 1. We controlled the degree of compliance of the condition by a *target ratio* parameter, defined as $t = \Delta_{max}/(avg^+ + avg^-)$. The condition is satisfied if and only if $t \in [0, 1]$, and smaller target-ratio values correspond to better fulfilment of the condition, and vice versa.

Given a desired target ratio, edge weights are generated as follows. First, all weights are drawn uniformly at random from a desired $[lb, ub]$ range. Then, the weights are adjusted in a two-step iterative fashion, until the desired target ratio is achieved: (i) the maximum gap Δ_{max} is fixed, the weights are changed for pairs that do not contribute to Δ_{max} so as to reflect a change in avg^+, avg^- ; (ii) avg^+, avg^- are fixed, Δ_{max} is updated by randomly modifying pairs that contribute to Δ_{max} . Finally, weight pairs are randomly assigned to the edges.

We compared the performance of Pivot (Algorithm 1 [4]) to one of the state-of-the-art algorithms achieving factor- $\mathcal{O}(\log |V|)$ guarantee on general graphs/ weights [6]. We dub the latter LP+R, alluding to the fact that it rounds the solution of a linear program. We evaluated correlation-clustering objective, number of output clusters, and runtimes of these algorithms.

Results. Figure 1 shows the quality (i.e., MIN-CC objective) of the produced clusterings, with the bottom-left insets reporting the ratio between the performance of Pivot and LP+R. Results refer to target ratios t varied from $[0, 3]$, with stepsize 0.1, and weights generated with $lb = 0, ub = 1$.

²Publicly available at <http://konect.cc/networks/>

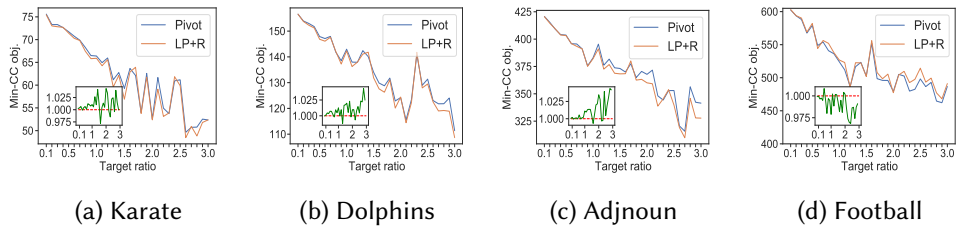


Figure 1: MIN-CC objective by varying the target ratio [3].

For each target ratio, all measurements correspond to averages over 10 weight-generation runs, and each of such runs corresponds to averages over 50 runs of the tested algorithms.

The main goal here is to have experimental evidence that a better fulfilment of our global condition leads to Pivot’s performance closer to LP+R’s one, and vice versa. Figure 1 confirms the above: in all datasets, Pivot performs more closely to LP+R as the target ratio gets smaller. In general, Pivot performs similarly to LP+R for $t \in [0, 1]$, while being outperformed for $t > 1$. This conforms with the theory: on these small graphs, factor-5 Pivot’s approximation is close to factor- $\mathcal{O}(\log |V|)$ LP+R’s approximation. Pivot achieves the best performance on *Football*, where it outperforms LP+R even if the condition is not met. This is motivated by *Football*’s higher clustering coefficient and average degree, which help Pivot sample vertices (and, thus, build clusters) in dense regions of the graph. This is confirmed by the number of clusters (Table 1-(left)): Pivot yields more clusters than LP+R on all datasets but *Football*.

Concerning execution times, Pivot runs in less than one second, and as expected, it is extremely faster than LP+R, whose runtimes are about 2 seconds in Karate, 37 in Dolphins, and above 770 in Adjnoun and Football.³ The inefficiency of LP+R further emphasizes the importance of our result in extending the applicability of faster algorithms like Pivot.

We complement this stage of evaluation by testing different graph densities, and for target ratios $t = 1$ (borderline satisfaction of our condition) and $t = 20$ (far fulfilment of the condition). Again, the results (not shown) meet the expectations: in terms of clustering quality, Pivot performs closely to or better than LP+R for $t = 1$, while the opposite happens for $t = 20$. Denser graphs correspond to better Pivot performance. This is again explained since higher densities favor better Pivot’s random choices. Runtimes are not affected by graph density. This is expected as well, as LP+R runtimes are dominated by the time spent in building and solving the linear program, which depends on the number of vertices only, whereas variations in the runtimes of Pivot cannot be observed due to the small size of the datasets at hand.

5. Application to fair clustering

An important exploitation of our theoretical results concerns the selection of features that lead edge weights to express the best tradeoff between an accurate representation of the objects and the suitability of the correlation-clustering weights to ensure approximation guarantees. Our global-weight-bounds condition can be an effective way to the achievement of this tradeoff, and it can be fulfilled more easily than local weight bounds (e.g., in case of probability constraint, it

³Experiments were carried out on the Cresco6 cluster (<https://www.eneagrid.enea.it>)

Table 1

Average clustering-sizes for various target ratios on the real-world graph datasets (left) and main characteristics of the relational datasets used in the fair clustering scenario (right).

| | 0.1 | | 0.5 | | 1 | | 3 | | #objs. | #attrs. | fairness-aware (sensitive) attributes |
|-----------------|-------|-------|-------|-------|-------|-------|-------|--------|--------|---------|--|
| | Pivot | LP+R | Pivot | LP+R | Pivot | LP+R | Pivot | LP+R | | | |
| <i>Karate</i> | 21.75 | 17.18 | 29.61 | 27.93 | 27.22 | 24.66 | 28.17 | 26.81 | 32 561 | 7/8 | race, sex, country, education, occupation, marital-status, workclass, relationship |
| <i>Dolphins</i> | 49.25 | 50.59 | 45.3 | 38.67 | 49.57 | 44.45 | 48.89 | 43.66 | 41 188 | 18/3 | job, marital-status, education |
| <i>Adjnoun</i> | 70.35 | 65.93 | 80.97 | 75.86 | 90.76 | 84.93 | 91.27 | 79.78 | 10 127 | 17/3 | gender, marital-status, education-level |
| <i>Football</i> | 64.43 | 84.91 | 77.14 | 96.43 | 68.35 | 78.72 | 90.87 | 100.31 | 649 | 28/5 | sex, male_edu, female_edu, male_job, female_job |

is hard to find a subset of features leading to positive-type and negative-type weights summing exactly to one for all the object pairs). We showcase this capability in a task of *fair clustering*.

Let \mathcal{X} be a set of objects defined over a set of attributes \mathcal{A} , which is assumed to be partitioned into two sets: \mathcal{A}^F , which contains *fairness-aware* or *sensitive* attributes (e.g., gender, race, religion), and \mathcal{A}^{-F} , which includes the remaining, *non-sensitive* attributes. We also assume that part of the attributes might be numerical, and the others as categorical; we will use superscripts N and C to distinguish the two types, therefore $\mathcal{A}^F = \mathcal{A}_N^F \cup \mathcal{A}_C^F$ and $\mathcal{A}^{-F} = \mathcal{A}_N^{-F} \cup \mathcal{A}_C^{-F}$.

We consider a twofold fair-clustering objective: cluster the objects such that (i) the intra-cluster similarity and the inter-cluster similarity are maximized and minimized, respectively, according to the non-sensitive attributes; (ii) the intra-cluster similarity and the inter-cluster similarity are minimized and maximized, respectively, according to the sensitive attributes. Pursuing this second objective would help distribute similar objects (in terms of sensitive attributes) across different clusters, thus helping the formation of diverse clusters. This is beneficial to ensure that the distribution of groups defined on sensitive attributes within each cluster approximates the distribution across the dataset.

The task of fair clustering can be mapped to a MIN-CC instance where the positive-type and negative-type weights, respectively, can be defined as follows:

$$w_{uv}^+ := \psi^+ \left(\alpha_N^{-F} \cdot \text{sim}_{\mathcal{A}_N^{-F}}(u, v) + (1 - \alpha_N^{-F}) \cdot \text{sim}_{\mathcal{A}_C^{-F}}(u, v) \right) \quad (2)$$

$$w_{uv}^- := \psi^- \left(\alpha_N^F \cdot \text{sim}_{\mathcal{A}_N^F}(u, v) + (1 - \alpha_N^F) \cdot \text{sim}_{\mathcal{A}_C^F}(u, v) \right) \quad (3)$$

where $\alpha_N^F = |\mathcal{A}_N^F| / (|\mathcal{A}_N^F| + |\mathcal{A}_C^F|)$ and $\alpha_N^{-F} = |\mathcal{A}_N^{-F}| / (|\mathcal{A}_N^{-F}| + |\mathcal{A}_C^{-F}|)$ are coefficients to weight similarities proportionally to the size of the involved set of attributes, $\psi^+ = \exp(|\mathcal{A}^F| / (|\mathcal{A}^F| + |\mathcal{A}^{-F}|)) - 1$ and $\psi^- = \exp(|\mathcal{A}^{-F}| / (|\mathcal{A}^F| + |\mathcal{A}^{-F}|)) - 1$ are smoothing factors to penalize correlation-clustering weights that are computed on a small number of attributes (which is usually the case for sensitive attributes, and hence negative-type weights), and $\text{sim}_S(\cdot)$ denotes any object similarity function defined over the subspace S of the attribute set.

Problem 2 (Attribute Selection for Fair Clustering). *Given a set of objects \mathcal{X} over the attribute sets \mathcal{A}^F , \mathcal{A}^{-F} , find maximal subsets $S^F \subseteq \mathcal{A}^F$ and $S^{-F} \subseteq \mathcal{A}^{-F}$, with $|S^F| \geq 1$, $|S^{-F}| \geq 1$, s.t. the weights computed by Eqs. (2)–(3) satisfy the global-weight-bounds condition in Th. 1.*

Heuristics. Our first proposal to solve Problem 2 is a greedy heuristic, dubbed Greedy, which iteratively removes the attribute that leads to the correlation-clustering weights with the lowest target ratio until our global condition is satisfied. This algorithm runs in $\mathcal{O}(|\mathcal{X}|^2 |\mathcal{A}|^2)$ time

Table 2

Fair clustering results. Values correspond to averages over the dataset-specific statistics (values under the column ‘orig.-weights Min-CC obj.’ were normalized for each dataset prior to the average calculation).

| | #it | target ratio | $\%(w^+ > w^-)$ | orig.-weights Min-CC obj. | avg. Eucl. fairness | avg. #clusts. | intra-clust \mathcal{A}^{-F} | intra-clust \mathcal{A}^F | inter-clust \mathcal{A}^{-F} | inter-clust \mathcal{A}^F | time (seconds) |
|---------|-------|--------------|-----------------|---------------------------|---------------------|---------------|--------------------------------|-----------------------------|--------------------------------|-----------------------------|----------------|
| initial | – | 1.289 | 95.735 | 0.182 | 0.046 | 25.8 | 0.611 | 0.537 | 0.376 | 0.142 | – |
| Hlv | 19.75 | 0.96 | 88.19 | 0.435 | 0.054 | 4.5 | 0.461 | 0.231 | 0.377 | 0.145 | 481.281 |
| Hlv_B | 16.75 | 0.905 | 82.752 | 0.507 | 0.093 | 510.5 | 0.761 | 0.705 | 0.409 | 0.141 | 460.475 |
| Hmv | 11.25 | 0.981 | 96.630 | 0.124 | 0.032 | 22.3 | 0.556 | 0.383 | 0.311 | 0.139 | 387.605 |
| Hmv_B | 10.25 | 0.967 | 94.722 | 0.264 | 0.054 | 239.3 | 0.732 | 0.673 | 0.398 | 0.149 | 346.156 |
| Hlv_BW | 15.0 | 0.96 | 82.985 | 0.880 | 0.129 | 777.3 | 0.883 | 0.850 | 0.407 | 0.147 | 378.958 |
| Hmv_SW | 11.0 | 0.955 | 96.447 | 0.085 | 0.019 | 3.5 | 0.493 | 0.279 | 0.293 | 0.136 | 447.854 |
| Greedy | 7.75 | 0.966 | 95.558 | 0.105 | 0.037 | 15.0 | 0.581 | 0.507 | 0.381 | 0.145 | 3324.521 |

since, at each iteration, for each candidate attribute to be removed $\mathcal{O}(|\mathcal{X}|^2)$ similarities are computed to quantify the decrease of the target ratio. We also devised other heuristics which, like Greedy, remove one attribute at time, but exploit some easy-to-compute proxy measures to select the attribute that avoid the pairwise similarity computation for each candidate attribute. The Hlv (resp. Hmv) heuristic removes the least (resp. most) variable attribute where the variability is measured through normalized entropy for categorical attributes and with variation coefficient (capped to 1 if above 1) for numerical features. Hlv_B and Hmv_B, like the previous two heuristics, remove the least and most variable attribute, respectively, but the selection is constrained to the biggest set of features among \mathcal{A}^F and \mathcal{A}^{-F} , in order to try to balance their size. Hlv_BW removes the least variable attribute from the set (\mathcal{A}^F or \mathcal{A}^{-F}) which induces the highest average similarity value using the current weights, whereas Hmv_SW selects the most variable attribute from the set which induces the lowest average similarity value using the current weights. Note that all these heuristics (but Greedy) run in $\mathcal{O}(|\mathcal{X}|^2|\mathcal{A}|)$ time.

Data and results. We considered 4 relational datasets: *CreditCardCustomers*, *Adult*, *Bank*, and *Student*.⁴ For each of them, Table 1 shows the number of objects, a pair of values corresponding to the count of non-sensitive and sensitive attributes, and a description of the latter.

Table 2 summarizes results achieved by each of the above heuristics, on the various datasets, according to the following criteria (columns from left to right): number of iterations at convergence, target ratio, percentage of pairs u, v having $w_{uv}^+ > w_{uv}^-$; also, computed w.r.t. the full attribute space are: value of the objective function, average Euclidean fairness [9], average number of clusters, intra-cluster and inter-cluster similarities according to either the subset of sensitive attributes or the subset of non-sensitive attributes, and running time. Euclidean and Jaccard similarity functions are used for numerical and categorical attributes, resp., and the overall similarity is obtained by linear combination analogously to Eqs. (2)–(3). Note that higher values correspond to better performance for \mathcal{A}^F -based intra-cluster and \mathcal{A}^{-F} -based inter-cluster similarities, while the opposite holds for the other two measures and the Euclidean fairness. The first row in each table refers to the initial, full-attribute-space status of the relational network, as a *baseline*, whereby the global-weight-bounds condition is not satisfied.

Hlv_BW and Hlv_B tend to produce solutions that correspond to the highest (i.e., worst) value of the objective function and of the clustering size; this should be ascribed to the fact

⁴<https://www.kaggle.com/sakshigoyal7/credit-card-customers>; <https://archive.ics.uci.edu/ml/index.php>

that both heuristics favor the selection of the least variable attributes. By contrast, Hmv_SW is the best performing in terms of objective function and Euclidean fairness. This method also tends to produce very few clusters. Note that, while a higher number of clusters is found to be coupled with a worsening of the objective function, the opposite does not hold in general; also, in contrast to the intuition that a higher percentage of pairs having $w^+ > w^-$ should favor the grouping into fewer clusters, we observed that the clustering sizes are not necessarily ordered as with the percentage ordering. As far as efficiency, Hmv_B mostly provides the best time performance. While being one order of magnitude slower, Greedy tends to converge in less iterations, as it indeed removes fewer attributes than the other methods; we found that in some cases (e.g., *Student*, *Adult*, results not shown), this allows Greedy for compensating its expected higher cost per iteration. Notably, each method lowers the initial target ratio below 1 so as to satisfy the global condition, and the per-dataset best-performing method (not shown) improves all intra-/inter-cluster similarities and Euclidean fairness w.r.t. the baseline.

6. Conclusions

We discussed a novel perspective in correlation clustering, which considers global weight bounds. A sufficient condition is defined to extend the range of validity of approximation guarantees beyond local weight bounds, such as the probability constraint. Experimental results, including a case study in fair clustering, put in evidence of the usefulness of our approach.

One interesting future direction we plan to investigate is to define the edge weights (bounds) based on relational properties of the objects under an uncertain data modeling framework [10].

References

- [1] N. Bansal, A. Blum, S. Chawla, Correlation clustering, *Mach. Learn.* 56 (2004) 89–113.
- [2] F. Bonchi, D. García-Soriano, E. Liberty, Correlation clustering: from theory to practice, in: *Proc. ACM KDD Conf.*, 2014, p. 1972.
- [3] D. Mandaglio, A. Tagarelli, F. Gullo, Correlation Clustering with Global Weight Bounds, in: *Proc. ECML-PKDD*, 2021, pp. 499–515. doi:10.1007/978-3-030-86520-7_31.
- [4] N. Ailon, M. Charikar, A. Newman, Aggregating inconsistent information: Ranking and clustering, *JACM* 55 (2008) 23:1–23:27.
- [5] R. Shamir, R. Sharan, D. Tsur, Cluster graph modification problems, *Discret. Appl. Math.* 144 (2004) 173–182.
- [6] M. Charikar, V. Guruswami, A. Wirth, Clustering with qualitative information, *JCSS* 71 (2005) 360–383.
- [7] E. D. Demaine, D. Emanuel, A. Fiat, N. Immerlica, Correlation clustering in general weighted graphs, *TCS* 361 (2006) 172–187.
- [8] G. J. Puleo, O. Milenkovic, Correlation clustering with constrained cluster sizes and extended weights bounds, *SIAM J. Optim.* 25 (2015) 1857–1872.
- [9] S. S. Abraham, D. P, S. S. Sundaram, Fairness in clustering with multiple sensitive attributes, in: *Proc. EDBT Conf.*, 2020, pp. 287–298.
- [10] F. Gullo, G. Ponti, A. Tagarelli, S. Greco, An information-theoretic approach to hierarchical clustering of uncertain data, *Inf. Sci.* 402 (2017) 199–215. doi:10.1016/j.ins.2017.03.030.