Correlation Clustering with Global Weight Bounds

Domenico Mandaglio[1], Andrea Tagarelli[1], Francesco Gullo[2]

[1]DIMES Dept., University of Calabria, Rende (CS), Italy; [2]UniCredit, Rome, Italy

ECML PKDD 2021 VIRTUAL 13-17 September

- We focus for the first time on **global weight bounds** for correlation clustering, focusing on its minimization objective.
- We identify a sufficient condition on input weights' aggregate functions to **extend the validity range of the approximation guarantees** of existing correlation-clustering algorithms beyond the probability constraint.
- We experimentally assess that our condition is an effective **indicator of the empirical performance** of existing probability-constraint-aware correlation-clustering algorithms.
- We showcase our results in a real-world scenario of **fair clustering.**

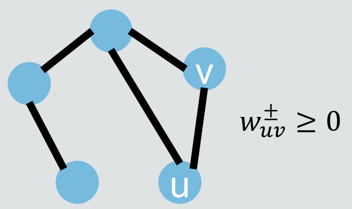
Min-Disagreement Correlation Clustering Problem

Given an undirected graph $G = (V, E)$, with vertex set V and edge set $E \subseteq V \times V$, and weights $w^+_{uv}, w^-_{uv} \in \mathbb{R}^+_0$ for all edges $(u, v) \in E$, find a clustering $\mathcal{C}: V \rightarrow \mathbb{N}^+$ that minimizes:

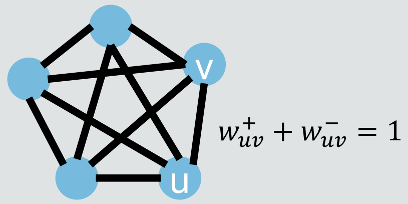
$$\sum_{\substack{(u,v)\in E \\ \mathcal{C}(u)=\mathcal{C}(v)}} w^-_{uv} + \sum_{\substack{(u,v)\in E \\ \mathcal{C}(u)\neq\mathcal{C}(v)}} w^+_{uv}$$

Any w^+_{uv} (resp. w^-_{uv}) weight expresses the benefit of clustering u and v together (resp. separately)

1. *G*eneral graph and general weights
   - Linear Programming + Rounding with $O(\log n)$ approximation guarantees



2. $E = \binom{V}{2}$ and $w^+_{uv} + w^-_{uv} = 1 \;\forall\, (u, v) \in E$
   - Pivot algorithm with (expected) 5-approximation guarantees and $O(|E|)$ time complexity

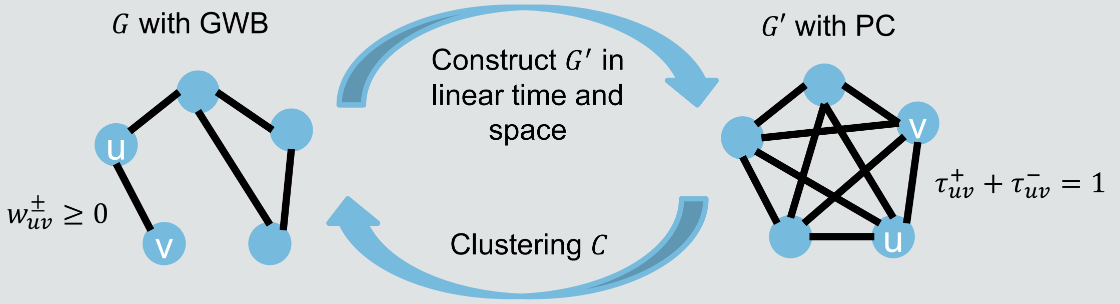


Can probability-constraint-aware approximation algorithms (e.g. Pivot) still achieve guarantees even if the probability constraint is not met?

Global Weight Bound (GWB) condition

$$\binom{|V|}{2}^{-1} \sum_{(u,v)\in E} w^+_{uv} + \binom{|V|}{2}^{-1} \sum_{(u,v)\in E} w^-_{uv} \geq \max_{(u,v)\in E} |w^+_{uv} - w^-_{uv}|$$

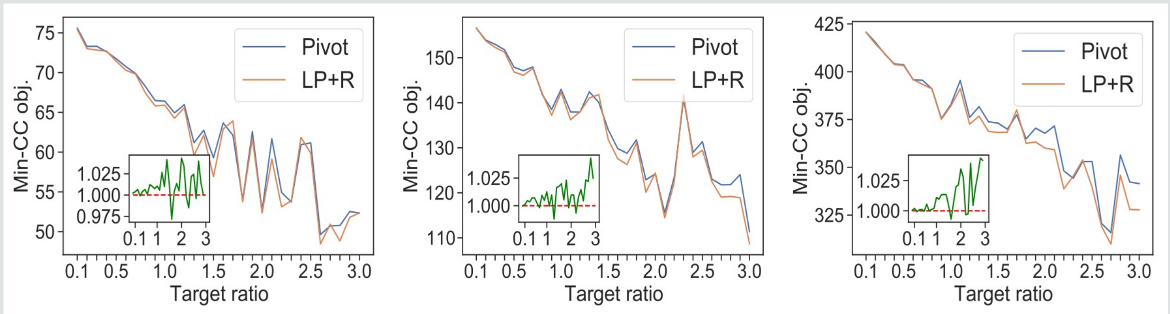
Strict approximation-preserving reduction:



An α-approximate clustering on G' is also α-approximate clustering on G too

- **Practical benefits:**
  - Extend the validity range of the approximation guarantees of algorithms for correlation clustering (Exp1)
  - Application to feature selection for fair clustering (Exp2)
- **Theoretical benefits:** enable better theoretical results on complex problems which exploit correlation clustering as a building block
- **Benefits for the research community:** brand new line of research

Exp1: Analysis of the global-weight-bounds condition



A better fulfilment of our GWB leads to Pivot's performance closer to the linear programming approach's one (LP+R, for short), and vice versa.

Exp2: Application to fair clustering

Data. Relational datasets describing a set of objects X defined over a set of attributes A (numerical or categorical) that can be divided into:

- *Fairness-aware* (or *sensitive*) attributes A^F
- *Non-sensitive* attributes $A^{\neg F}$

Fair clustering objective:

- ***non-sensitive attributes***: minimize the inter-cluster similarities and maximize the intra-cluster similarities
- ***sensitive attributes***: minimize the intra-cluster similarities and maximize the inter-cluster similarities

Mapping to Correlation Clustering instance:

$$w^+_{uv} := \varphi^+\left(\alpha^{\neg F}_N \cdot sim_{A^{\neg F}_N}(u, v) + \left(1 - \alpha^{\neg F}_N\right) \cdot sim_{A^{\neg F}_C}(u, v)\right)$$

$$w^-_{uv} := \varphi^-\left(\alpha^F_N \cdot sim_{A^F_N}(u, v) + (1 - \alpha^F_N) \cdot sim_{A^F_C}(u, v)\right)$$

$$\alpha^F_N = \frac{|A^F_N|}{|A^F_N| + |A^F_C|}, \alpha^{\neg F}_N = \frac{|A^{\neg F}_N|}{|A^{\neg F}_N| + |A^{\neg F}_C|}, \varphi^+ = \exp\left(\frac{|A^F|}{|A^F| + |A^{\neg F}|} - 1\right), \varphi^- = \exp\left(\frac{|A^{\neg F}|}{|A^F| + |A^{\neg F}|} - 1\right)$$

Attribute selection for fair clustering. Given a set of objects X defined over the attribute sets A^F and $A^{\neg F}$, find maximal subsets $S_F \subseteq A^F$ and $S_{\neg F} \subseteq A^{\neg F}$, with $|S_F| \geq 1$ and $|S_{\neg F}| \geq 1$, s.t. the above correlation-clustering weights satisfy the GWB condition.

	#it	target ratio	$\%(w^+ > w^-)$	orig.-weights Min-CC obj.	avg. Eucl. fairness	avg. #clusts.	intra-clust $\mathcal{A}^{\neg F}$	intra-clust $\mathcal{A}^F$	inter-clust $\mathcal{A}^{\neg F}$	inter-clust $\mathcal{A}^F$	time (seconds)
						Adult					
initial	–	1.086	90.34	1.1915E+08	0.082	77	0.699	0.672	0.378	0.181	–
Hlv	12	0.986	93.19	1.122659E+08	**0.031**	9	0.465	0.326	0.347	0.194	545.249
Hlv_B	12	0.765	78.09	1.119757E+08	0.039	69	0.608	0.547	0.375	0.184	529.674
Hmv	5	0.974	90.83	1.21187E+08	0.094	79	0.689	0.687	0.373	**0.203**	220.056
Hmv_B	4	0.936	87.39	1.25516E+08	0.109	905	0.963	0.96	0.377	0.199	**178.813**
Hlv_BW	5	0.963	83.17	1.343503E+08	0.152	1479	**0.969**	0.964	0.384	0.199	217.333
Hmv_SW	9	0.926	91.41	1.159874E+08	0.037	5	0.451	**0.308**	**0.329**	0.195	380.875
Greedy	2	0.967	92.36	**1.094787E+08**	0.036	32	0.668	0.654	0.361	0.195	595.610

The GWB condition can help to define weights so as to account for both an effective representation of the semantics underlying objects' features, and the peculiarities that make the downstream correlation-clustering algorithm effective.

[1]{d.mandaglio, tagarelli}@dimes.unical.it; [2]gullof@acm.org