UNIVERSITÀ DELLA CALABRIA
DIPARTIMENTO DI INGEGNERIA INFORMATICA, MODELLISTICA, ELETTRONICA E SISTEMISTICA
DIMES

UniCredit

*The views and opinions expressed in this paper are those of the author and do not necessarily reflect the official policy or position of the UniCredit group.

Correlation Clustering with Global Weight Bounds

Domenico Mandaglio, Andrea Tagarelli, Francesco Gullo*

DIMES – Univ. Calabria
Rende (CS), Italy

DIMES – Univ. Calabria
Rende (CS), Italy

UniCredit
Rome, Italy

Outline

- Background: Correlation Clustering with ***local*** weight bounds
- This work: Correlation Clustering with ***global*** weight bounds
- Theoretical results and algorithms
- Experimental results
- Conclusions & Future Work

Min-Disagreement Correlation Clustering (Min-CC)

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 $\mathcal{C}: V \to \mathbb{N}^+$ 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)

Min-Disagreement Correlation Clustering (Min-CC)

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 $\mathcal{C}: V \to \mathbb{N}^{+}$ 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)

- Min-CC is **NP**-Hard
- **APX**-Hard even for complete graphs and edge weights $(w_{uv}^{+},\ w_{uv}^{-}) \in \{(0,1), (1,0)\}$
- For general graphs and general weights the best known approximation factor is $O(\log(|V|))$, on rounding the solution to a large linear program[1] (with a number of $\Omega(|V|^3)$ constraints)

1. Charikar Moses, Venkatesan Guruswami, and Anthony Wirth. "Clustering with qualitative information." Journal of Computer and System Sciences 71.3 (2005): 360-383.

Special case for Min-CC

- Complete graph: $E = \binom{V}{2}$
- Probability constraint (PC): $w_{uv}^{+} + w_{uv}^{-} = 1 \; \forall \, (u, v) \in E$

Special case for Min-CC

- Complete graph: $E = \binom{V}{2}$
- Probability constraint (PC): $w_{uv}^{+} + w_{uv}^{-} = 1 \; \forall \, (u, v) \in E$

Pivot algorithm[2]

- Pick a node u uniformly at random
- Build a cluster upon u together with its neighbor similar nodes that are still unclustered
- Remove the built cluster from the graph
- Repeat until the graph is empty

Properties of Pivot:

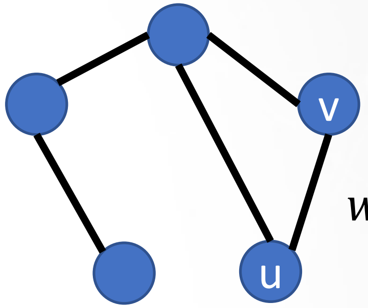
- (expected) 5-approximation guarantee
- Efficiency: $O(|E|)$ time complexity
- Easy-to-implement

2. Ailon Nir, Moses Charikar, and Alantha Newman. "Aggregating inconsistent information: ranking and clustering." Journal of the ACM (JACM) 55.5 (2008): 1-27.

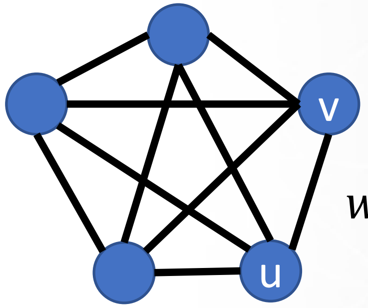
General vs Constrained Min-CC instances

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

2. Complete graph and $w_{uv}^{+} + w_{uv}^{-} = 1 \ \forall \ (u, v) \in E$
   - Pivot algorithm with constant-factor approximation guarantees



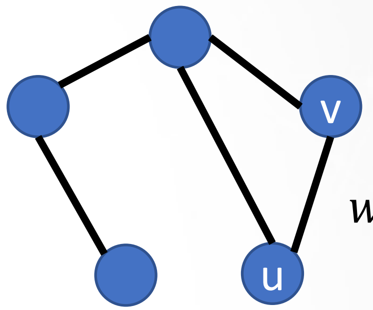
$w_{uv}^{\pm} \geq 0$



$w_{uv}^{+} + w_{uv}^{-} = 1$

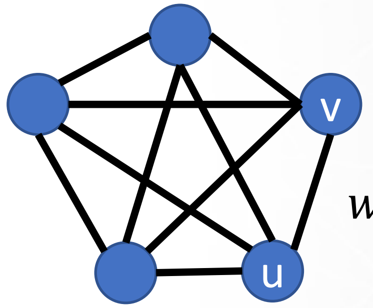
General vs Constrained Min-CC instances

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



$w_{uv}^{\pm} \geq 0$

2. Complete graph and $w_{uv}^{+} + w_{uv}^{-} = 1 \; \forall \, (u, v) \in E$
   - Pivot algorithm with constant-factor approximation guarantees



$w_{uv}^{+} + w_{uv}^{-} = 1$

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

Min-CC with Global Weight Bounds: Theoretical Results and Algorithms

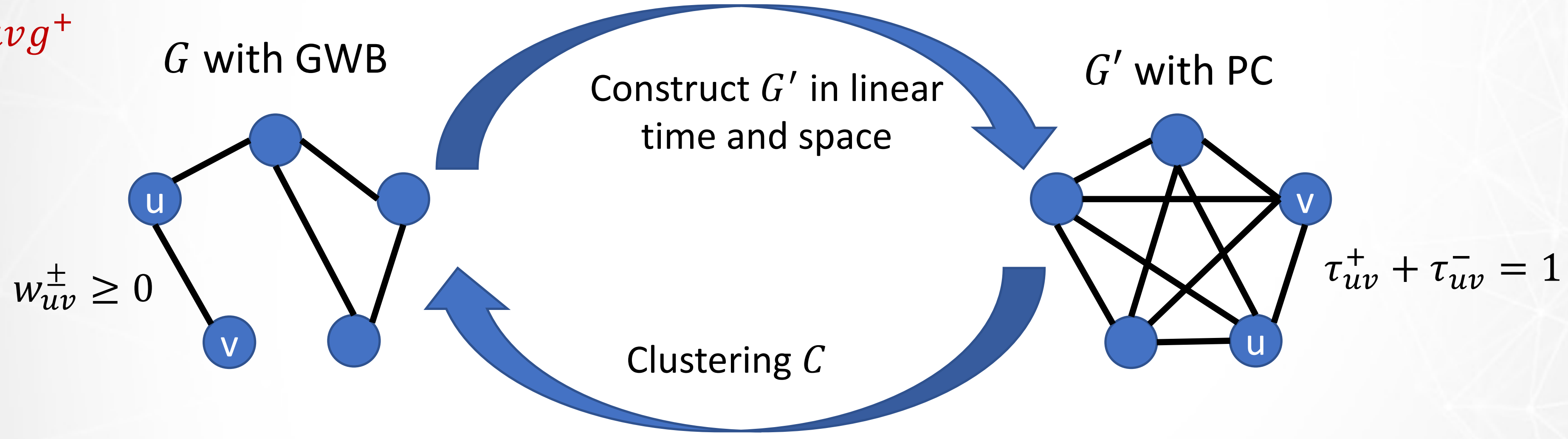
Global Weight Bound (GWB):

$$\underbrace{\binom{|V|}{2}^{-1} \sum_{(u,v)\in E} w_{uv}^{+}}_{avg^{+}} + \overbrace{\binom{|V|}{2}^{-1} \sum_{(u,v)\in E} w_{uv}^{-}}^{avg^{-}} \geq \overbrace{\max_{(u,v)\in E} |w_{uv}^{+} - w_{uv}^{-}|}^{\Delta_{max}}$$

Min-CC with Global Weight Bounds: Theoretical Results and Algorithms

Global Weight Bound (GWB):

$$\underbrace{\binom{|V|}{2}^{-1} \sum_{(u,v)\in E} w_{uv}^{+}}_{avg^{+}} + \overbrace{\binom{|V|}{2}^{-1} \sum_{(u,v)\in E} w_{uv}^{-}}^{avg^{-}} \geq \overbrace{\max_{(u,v)\in E} |w_{uv}^{+} - w_{uv}^{-}|}^{\Delta_{max}}$$



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

Min-CC with Global Weight Bounds: Theoretical Results and Algorithms

---

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 $\tau_{uv}^+, \tau_{uv}^-, \forall u, v \in V$, as in Equation (3) (using M, γ defined in Step 1)

3: $\mathcal{C} \leftarrow$ run A on Min-pc-CC instance $\langle G' = (V, V \times V), \{\tau_e^+, \tau_e^-\}_{e \in V \times V} \rangle$

Corollary: Let I be a Min-CC instance satisfying the GWB, and A be an α-approximation algorithm for Min-CC with PC. GlobalCC on input $< I, A >$ achieves factor-α guarantee on I.

Benefits of our result

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

Exp1: Analysis of the global-weight-bounds condition

Data: 4 real-world graphs augmented with artificially-generated edge weights, to test different levels of fulfilment (controlled by the parameter *target ratio*) of our global-weight-bounds (GWB) condition.

$$\Delta_{max}/(avg^{+} + avg^{-}) \leq 1$$

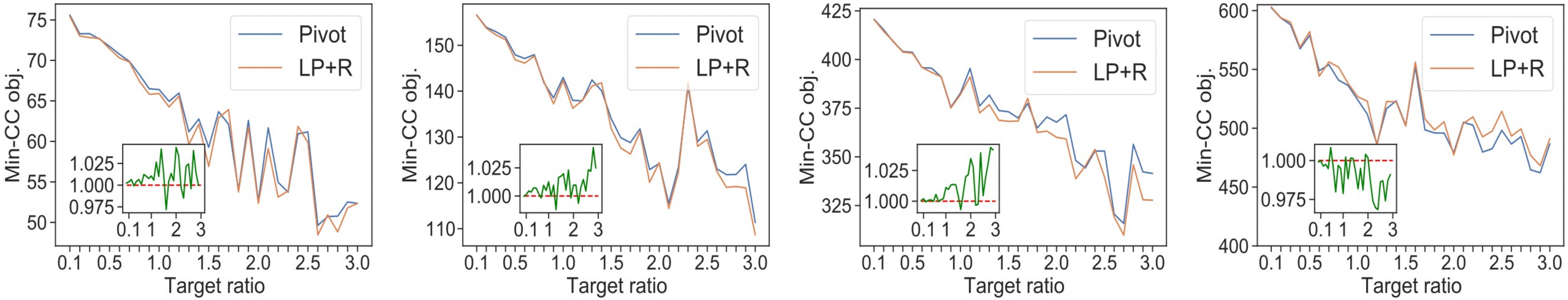
$$\Updownarrow$$

GWB: $avg^{+} + avg^{-} \geq \Delta_{max}$

	$\lvert V\rvert$	$\lvert E\rvert$	den.	a_deg	a_pl	diam	cc
Karate	34	78	0.14	4.59	2.41	5	0.26
Dolphins	62	159	0.08	5.13	3.36	8	0.31
Adjnoun	112	425	0.07	7.59	2.54	5	0.16
Football	115	613	0.09	10.66	2.51	4	0.41

Goal: show that a better fulfilment of the GWB corresponds to better performance (in terms of Min-CC objective) of Pivot with respect to the LP algorithms, and vice versa.

Exp1: Analysis of the global-weight-bounds condition



(a) Karate (b) Dolphins (c) Adjnoun (d) Football

Fig. 1: MIN-CC objective by varying the target ratio.

A better fulfilment of our GWB leads to Pivot's performance closer to the linear programming approach's one[1] (LP+R, for short), and vice versa

1. Charikar Moses, Venkatesan Guruswami, and Anthony Wirth. "Clustering with qualitative information." Journal of Computer and System Sciences 71.3 (2005): 360-383.

Exp1: Analysis of the global-weight-bounds condition

Table 2: Running times (left) and avg. clustering-sizes for various target ratios (right).

	Pivot (secs.)	LP+R (secs.)
Karate	< 1	1.9
Dolphins	< 1	36.58
Adjnoun	< 1	775.4
Football	< 1	819.8

	0.1		0.5		1		2		3	
	Pivot	LP+R	Pivot	LP+R	Pivot	LP+R	Pivot	LP+R	Pivot	LP+R
Karate	21.75	17.18	29.61	27.93	27.22	24.66	25.55	23.82	28.17	26.81
Dolphins	49.25	50.59	45.3	38.67	49.57	44.45	47.91	48.05	48.89	43.66
Adjnoun	70.35	65.93	80.97	75.86	90.76	84.93	85.83	70.41	91.27	79.78
Football	64.43	84.91	77.14	96.43	68.35	78.72	78.65	85.31	90.87	100.31

- Pivot is faster than LP+R
- Pivot yields more clusters than LP+R on all datasets but Football

Exp2: Application to fair clustering

Data: 4 real-world 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 $A^{\neg F}$

	#objs.	#attrs.	fairness-aware (sensitive) attributes
Adult	32 561	7/8	race, sex, country, education, occupation, marital-status, workclass, relationship
Bank	41 188	18/3	job, marital-status, education
Credit	10 127	17/3	gender, marital-status, education-level
Student	649	28/5	sex, male_edu, female_edu, male_job, female_job

Exp2: Application to fair clustering

Fair clustering objective:

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

Fairness requirement: distribute similar objects (in terms of sensitive attributes) across different clusters, thus helping the formation of diverse clusters.

Exp2: Application to fair clustering

Mapping to Min-CC instance:

$$w_{uv}^{+} := \varphi^{+}\left(\alpha_N^{\neg F} \cdot sim_{A_N^{\neg F}}(u,v) + \left(1 - \alpha_N^{\neg F}\right) \cdot sim_{A_C^{\neg F}}(u,v)\right)$$

$$w_{uv}^{-} := \varphi^{-}\left(\alpha_N^{F} \cdot sim_{A_N^{F}}(u,v) + (1 - \alpha_N^{F}) \cdot sim_{A_C^{F}}(u,v)\right)$$

$$\alpha_N^F = \frac{|A_N^F|}{|A_N^F| + |A_C^F|}, \alpha_N^{\neg F} = \frac{|A_N^{\neg F}|}{|A_N^{\neg F}| + |A_C^{\neg F}|}, \varphi^{+} = \exp\left(\frac{|A^F|}{|A^F| + |A^{\neg F}|} - 1\right), \varphi^{-} = \exp\left(\frac{|A^{\neg F}|}{|A^F| + |A^{\neg F}|} - 1\right)$$

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 global-weight-bounds condition.

Exp2: Application to fair clustering

Table 3: Fair clustering results.

	#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
						Bank					
initial	–	1.612	98.84	7.738171E+07	0.019	9	0.593	0.466	0.413	0.083	–
Hlv	19	0.95	99.88	7.063441E+07	0.001	3	0.52	0.209	0.368	**0.082**	1289.785
Hlv_B	16	0.906	97.19	8.489668E+07	0.038	752	0.859	0.818	0.456	0.077	1223.205
Hmv	17	0.972	100.0	**7.032421E+07**	**0.0**	2	0.497	**0.136**	**0.151**	0.03	1254.341
Hmv_B	16	0.981	97.19	8.250374E+07	0.032	35	0.775	0.665	0.451	0.079	**1143.517**
Hlv_BW	17	0.984	92.87	1.163447E+08	0.095	1048	**0.997**	0.996	0.444	0.076	1212.091
Hmv_SW	17	0.972	100.0	**7.032421E+07**	**0.0**	2	0.497	**0.136**	**0.151**	0.03	1336.888
Greedy	13	0.981	99.57	7.240143E+07	0.006	3	0.508	0.371	0.381	0.076	11978.472

Each method decreases the initial target ratio below 1 so as to satisfy the global condition, and the per-dataset best-performing method improves all intra-/inter-cluster similarities and Euclidean fairness w.r.t. the baseline.

Conclusion & Future Work

Summary:

- We studied for the first time global weight bounds in correlation clustering
- We derived a sufficient condition to extend the range of validity of approximation guarantees beyond local weight bounds, such as the probability constraint

Future Work:

- extending our results to other constraints (e.g., triangle inequality)
- studying the by-product problem of feature selection guided by our condition