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: from Local to Global Constraints

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 $\mathcal{C}: V \rightarrow \mathbb{N}^{+}$ that minimizes:

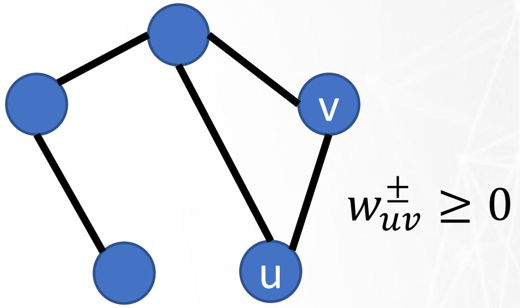
$$\sum_{\substack{(u,v)\in E \\ \mathcal{C}(u)=\mathcal{C}(v)}} w_{uv}^{-} + \sum_{\substack{(u,v)\in E \\ \mathcal{C}(u)\neq\mathcal{C}(v)}} w_{uv}^{+}$$

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

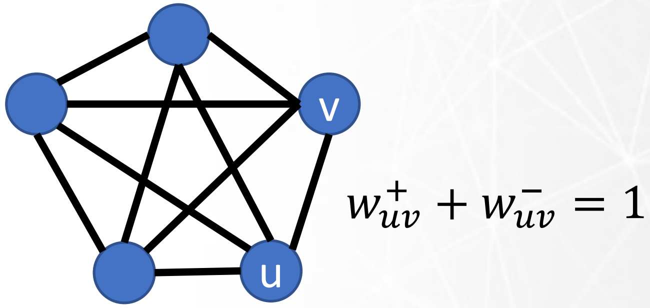
- Min-CC is **NP**-Hard
- **APX**-Hard even for complete graphs and edge weights $(w_{uv}^{+},\ w_{uv}^{-}) \in \{(0,1), (1,0)\}$

Approximation Algorithms: General vs Constrained Min-CC instances

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



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

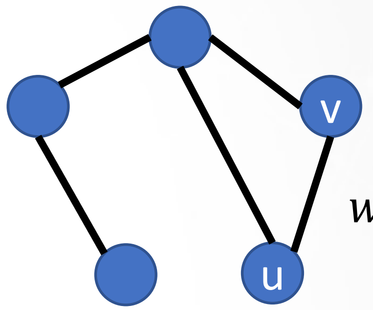


1. Charikar Moses, Venkatesan Guruswami, and Anthony Wirth. "Clustering with qualitative information." Journal of Computer and System Sciences 71.3 (2005): 360-383.
2. Ailon Nir, Moses Charikar, and Alantha Newman. "Aggregating inconsistent information: ranking and clustering." Journal of the ACM (JACM) 55.5 (2008): 1-27.

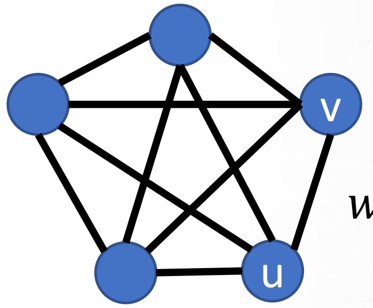
Approximation Algorithms: General vs Constrained Min-CC instances

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

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



$w^{\pm}_{uv} \geq 0$



$w^{+}_{uv} + w^{-}_{uv} = 1$

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

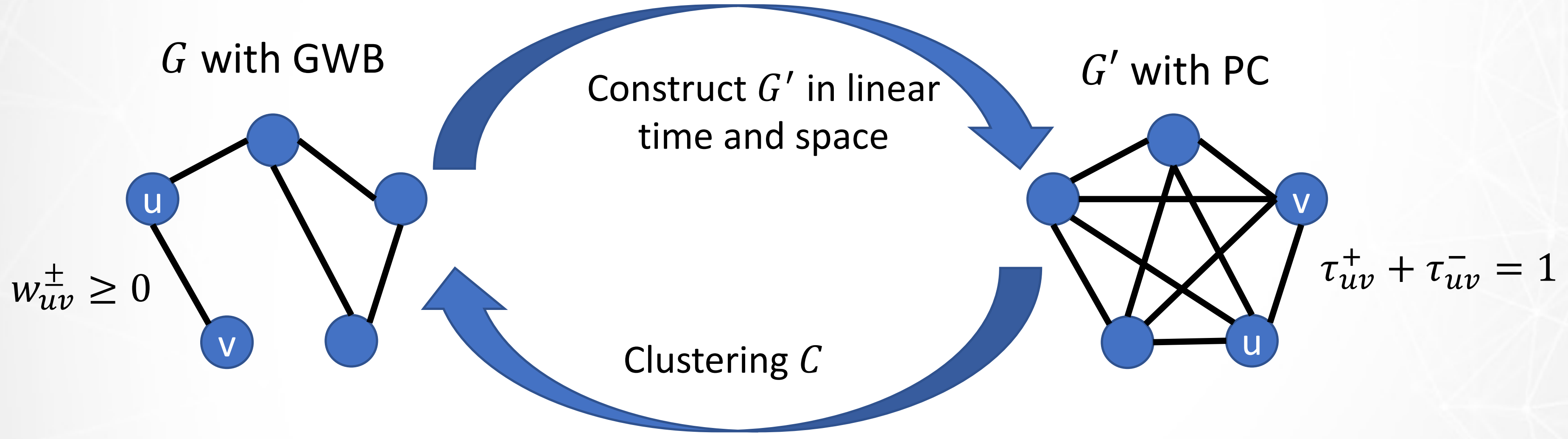
Global Weight Bound (GWB):

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

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

Global Weight Bound (GWB):

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



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

Benefits of our result

- **Practical benefits:**
  - Extend the validity range of the approximation guarantees of algorithms for Min-CC (e.g. Pivot)
  - Application to feature selection for fair clustering (Next slides)
- **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

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:

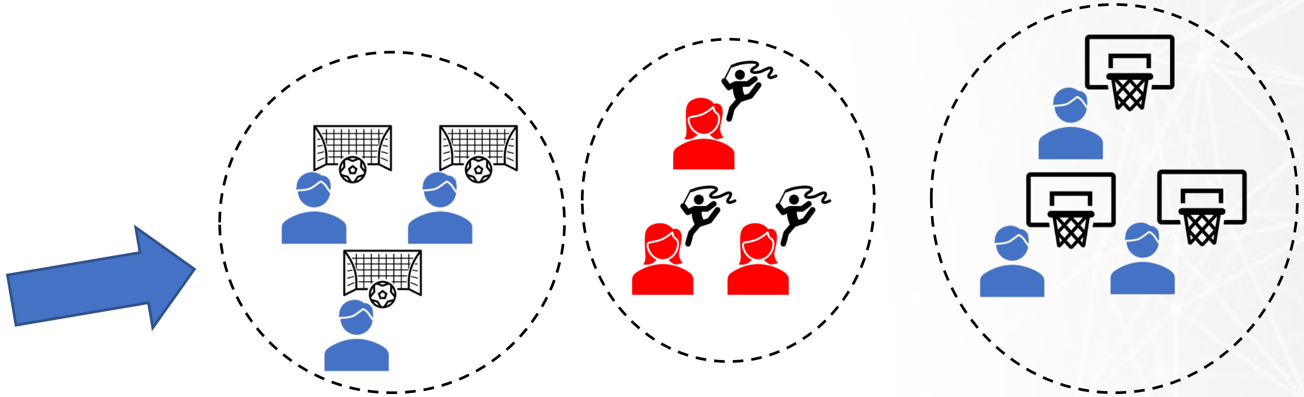
- *Fairness-aware* (or *sensitive*) attributes A^F
- *Non-sensitive* attributes $A^{\neg F}$

	#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

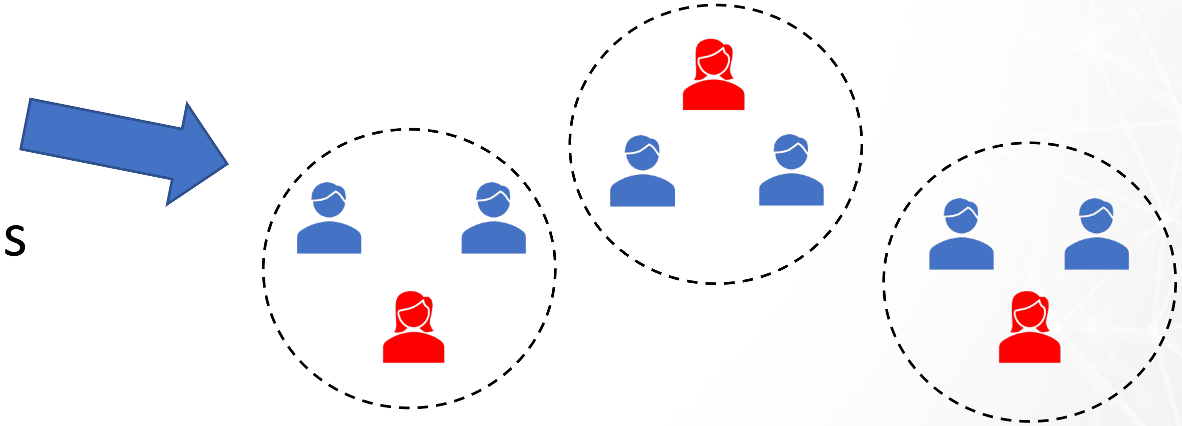
Application to fair clustering

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



Application to fair clustering

Mapping to Min-CC instance:

$$\mathrm{G} = (\mathrm{V} = \mathrm{X}\,, E = X \times X) \qquad w_{uv}^{+} \coloneqq\propto sim_{A^{\neg F}}(u, v) \qquad w_{uv}^{-} \coloneqq\propto sim_{A^{F}}(u, v)$$

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.

Application to fair clustering

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 $\mathcal{A}^{\neg F}$	intra-clust $\mathcal{A}^{F}$	inter-clust $\mathcal{A}^{\neg F}$	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

Each method finally finds two subsets of attributes 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 ('initial' in the Table).

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

Thanks for your attention!
Questions?

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.

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

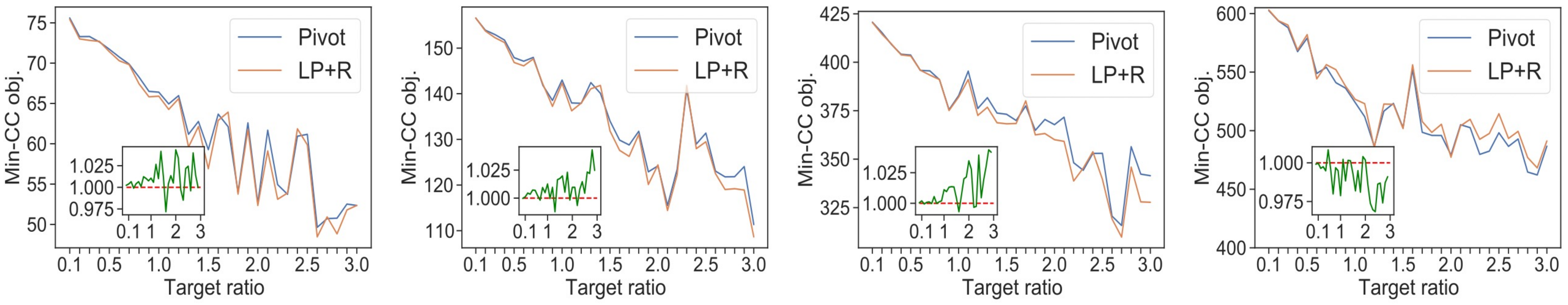
$$\Updownarrow$$

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

Mapping to Min-CC instance:

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

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

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

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.