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.

DS 2022

25th International Conference on Discovery Science
Montpellier, France - 10-12 October 2022

When Correlation Clustering Meets Fairness Constraints

Francesco Gullo, Lucio La Cava, Domenico Mandaglio, Andrea Tagarelli

UniCredit Rome, Italy

DIMES - Univ. Calabria Rende (CS) Italy

DIMES - Univ. Calabria Rende (CS) Italy

DIMES - Univ. Calabria Rende (CS) Italy

Today’s Menu

- Intro to the context
- Background on Correlation Clustering
- The Fair-CC Problem
- Proposed approach
- Fairness-aware evaluation metrics
- Experimental methodology and results
- Conclusions and Future Work

Introduction

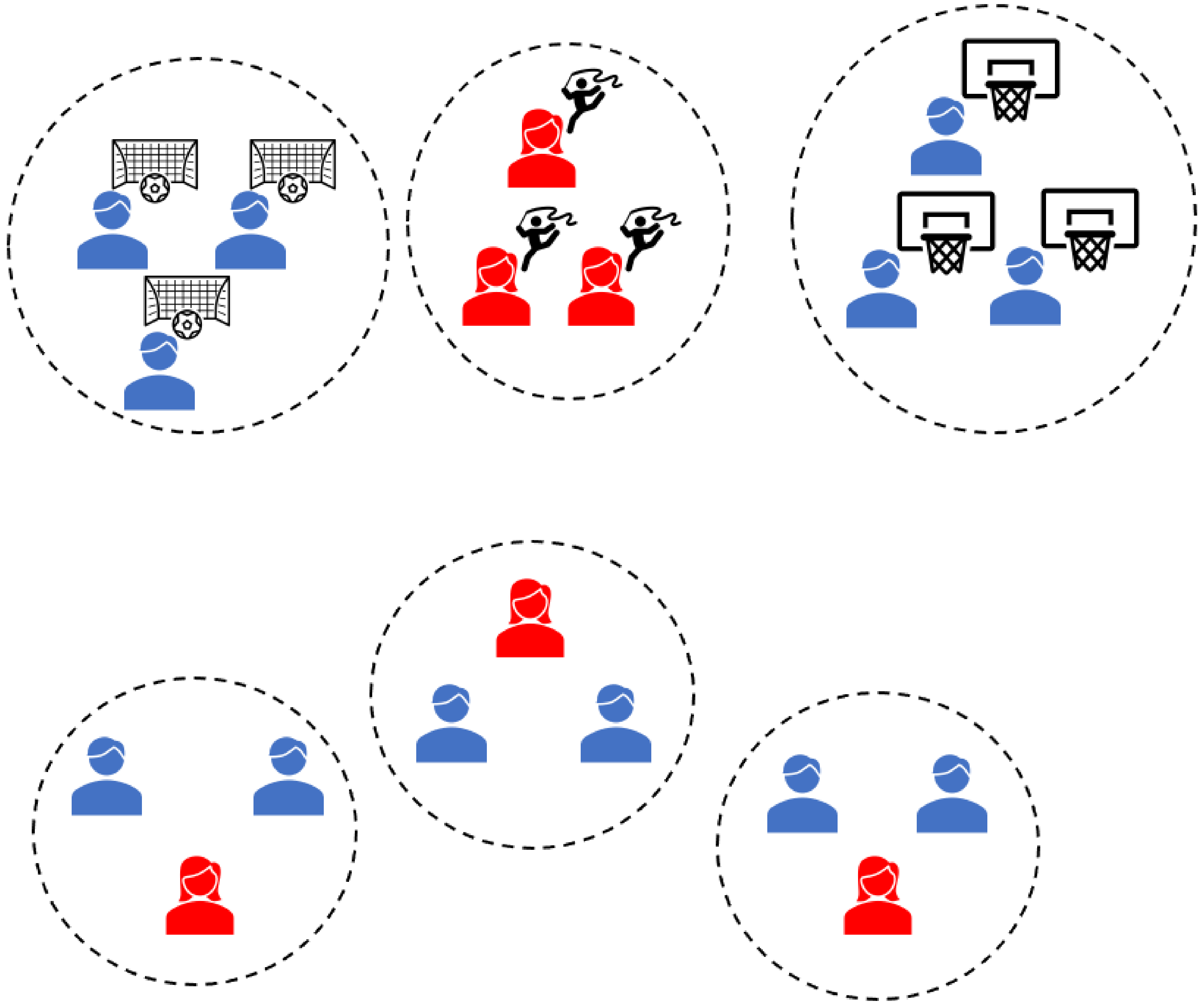
- Machine Learning (ML) systems achieved decision-making power in our lives (shall we entrust them?)
- Input data is often (intrinsically) biased
- ML algorithms must avoid amplifying input data bias
- Disparate impact must be removed
  - *no group of individuals should (even indirectly) be discriminated by a decision-making system* [1]

[1] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proc. ACM KDD Conf. pp. 259–268 (2015)

The Fair Clustering Problem

Clustering a set of data objects s.t.:

- Similar objects are assigned to the same cluster, whereas dissimilar objects are assigned to different clusters
- Clusters should not be dominated by a specific type of sensitive data class (e.g., people having the same sex)



Can we tackle this problem through a
correlation clustering framework?

Min-Disagreement Correlation Clustering (MIN-CC)

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

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

where w_{uv}^{+}, resp. w_{uv}^{-}, denote the benefit of clustering u and v together, resp. separately.

Problem Statement - Notation

Let $\mathcal{X} = \{X_1, \cdots, X_n\}$ be a set of n objects defined over a set of attributes $\mathcal{A}$ divided into two sets:

- $\mathcal{A}^F$ containing *fairness-aware* (or *sensitive*) attributes (e.g., those identifying sex, race, religion, relationship status in a citizen database);
- $\mathcal{A}^{\neg F}$ containing *non-sensitive* attributes (e.g., user preferences).

Both can include numerical (N) and categorical (C) attributes:

$$\mathcal{A}^F = \mathcal{A}_N^F \cup \mathcal{A}_C^F, \qquad \mathcal{A}^{\neg F} = \mathcal{A}_N^{\neg F} \cup \mathcal{A}_C^{\neg F}$$

Problem Statement - Fair-CC

Given a set of objects $\mathcal{X}$, two sets of attributes $\mathcal{A}^F$ and $\mathcal{A}^{\neg F}$, and an object similarity function $sim_S(\cdot)$ defined over the subspace S of the attribute set, find a clustering $\mathcal{C}^*$ to minimize:

$$\sum_{u,v \in \mathcal{X}, \mathcal{C}(u)=\mathcal{C}(v)} sim_{\mathcal{A}^F}(u,v) + \sum_{u,v \in \mathcal{X}, \mathcal{C}(u)\neq\mathcal{C}(v)} sim_{\mathcal{A}^{\neg F}}(u,v)$$

This corresponds to solving a complete Min-CC instance:

- The set of vertices corresponds to the objects in $\mathcal{X}$ and,
- For each pair of vertices u and v, the positive-type (resp. negative-type) correlation-clustering weight corresponds to the similarity score between the two vertices according to the non-sensitive (resp. sensitive) attributes.

Utility functions

$$sim_{\mathcal{A}^{\neg F}}(u,v) := \psi^{+}(\alpha_N^{\neg F} \cdot sim_{\mathcal{A}_N^{\neg F}}(u,v) + (1-\alpha_N^{\neg F}) \cdot sim_{\mathcal{A}_C^{\neg F}}(u,v))$$
$$sim_{\mathcal{A}^{F}}(u,v) := \psi^{-}(\alpha_N^{F} \cdot sim_{\mathcal{A}_N^{F}}(u,v) + (1-\alpha_N^{F}) \cdot sim_{\mathcal{A}_C^{F}}(u,v))$$

Similarity according to the set of non-sensitive and sensitive attributes

$$\alpha_N^F = |\mathcal{A}_N^F| / (|\mathcal{A}_N^F| + |\mathcal{A}_C^F|)$$
$$\alpha_N^{\neg F} = |\mathcal{A}_N^{\neg F}| / (|\mathcal{A}_N^{\neg F}| + |\mathcal{A}_C^{\neg F}|)$$

Weight similarities proportionally to the number of involved attributes

$$\psi^{+} = exp(|\mathcal{A}^F| / (|\mathcal{A}^F| + |\mathcal{A}^{\neg F}|) - 1)$$
$$\psi^{-} = exp(|\mathcal{A}^{\neg F}| / (|\mathcal{A}^F| + |\mathcal{A}^{\neg F}|) - 1)$$

Smoothing factors to penalize weights that are computed on a small number of attributes

Solving Fair-CC

The CC-Bounds algorithm: [2]

Input: Set of objects $\mathcal{X}$, sensitive attributes $\mathcal{A}^F$, non-sensitive attributes $\mathcal{A}^{\neg F}$, Min-CC algorithm A

Output: Clustering $\mathcal{C}$ of $\mathcal{X}$

1. Compute $sim_{\mathcal{A}^{\neg F}}(u, v), sim_{\mathcal{A}^F}(u, v)\ \forall u, v \in \mathcal{X}$
2. Build the instance $I = \langle G = (\mathcal{X}, \mathcal{X} \times \mathcal{X}), \{sim_{\mathcal{A}^{\neg F}}(u, v), sim_{\mathcal{A}^F}(u, v)\}_{u,v \in \mathcal{X} \times \mathcal{X}} \rangle$
3. $\mathcal{C} \leftarrow$ run A on I

[2] Mandaglio, D., Tagarelli, A., Gullo, F.: Correlation clustering with global weight bounds. In: Proc. ECML-PKDD Conf. pp. 499–515 (2021)

Theoretical remarks

Let $T_A(\mathcal{X})$ the running time of the algorithm A on the set of data objects $\mathcal{X}$

➡ The time complexity of CCBounds is $\mathcal{O}(|\mathcal{X}|^2|\mathcal{A}| + T_A(\mathcal{X}))$

- Compute similarities over $\mathcal{A}$ attributes, for each pair of objects in $\mathcal{X}$, then solve the resulting Min-CC instance through A

➡ The space complexity of CC-Bounds is $\mathcal{O}(|\mathcal{X}|^2)$

- In-memory similarity storing

The Min-CC algorithm A used in CC-Bounds is the one proposed in [3], as it proposes constant-factor approximation guarantees (under certain conditions), s.t. $T_A(\mathcal{X}) = \mathcal{O}(|\mathcal{X}|^2)$.

✓ The time complexity of CCBounds become $\mathcal{O}(|\mathcal{X}|^2|\mathcal{A}|)$.

[3] Ailon, N., Charikar, M., Newman, A.: Aggregating inconsistent information: ranking and clustering. In: Proc. ACM STOC Symp. pp. 684–693 (2005)

Theorem 1 [2]

If the condition

$$\binom{|\mathcal{X}|}{2}^{-1}\left(sim_{\mathcal{A}^{\neg F}}(u,v)+sim_{\mathcal{A}^{F}}(u,v)\right)$$
$$\geq \max_{u,v\in\mathcal{X}}\left|sim_{\mathcal{A}^{\neg F}}(u,v)-sim_{\mathcal{A}^{F}}(u,v)\right|$$

holds on the similarity scores and the oracle A is an α-approximation algorithm for Min-CC, CCBounds is α-approximation algorithm for Fair-CC.

[2] Mandaglio, D., Tagarelli, A., Gullo, F.: Correlation clustering with global weight bounds. In: Proc. ECML-PKDD Conf. pp. 499–515 (2021)

Evaluating Fairness

Focus on algorithm-independent evaluation metrics following a *group-level* approach under the *disparate impact* doctrine [4]

$$balance(\mathcal{C}) \overset{[5,\,6]}{=} \min_{C \in \mathcal{C}, b \in [m]} \min\{R_{C,b}, \frac{1}{R_{C,b}}\} \in [0,1]$$

$$AE_A(\mathcal{C}) \overset{[7]}{=} \frac{\sum_{C \in \mathcal{C}} |C| \times ED(C_A, \mathcal{X}_A)}{\sum_{C \in \mathcal{C}} |C|}$$

[4] Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proc. ACM KDD Conf. pp. 259–268 (2015)

[5] Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. In: Proc. NIPS Conf. pp. 5029–5037 (2017)

[6] Bera, S.K., Chakrabarty, D., Flores, N., Negahbani, M.: Fair algorithms for clustering. In: Proc. NIPS Conf. pp. 4955–4966 (2019)

[7] Abraham, S.S., P, D., Sundaram, S.S.: Fairness in clustering with multiple sensitive attributes. In: Proc. EDBT Conf. pp. 287–298 (2020)

Competing methods

- *Fair Clustering through Fairlets* [5]
- *HST-based Fair Clustering* [8]
- *Fair Correlation Clustering* [9]

- Based on *fairlets decomposition* (direct or via correlation clustering)
- The first two can just handle a single sensitive attribute

[5] Chierichetti, F., Kumar, R., Lattanzi, S., Vassilvitskii, S.: Fair clustering through fairlets. In: Proc. NIPS Conf. pp. 5029–5037 (2017)
[8] Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., Wagner, T.: Scalable fair clustering. In: Proc. ICML Conf. pp. 405–413 (2019)
[9] Ahmadian, S., Epasto, A., Kumar, R., Mahdian, M.: Fair correlation clustering. In: Proc. AISTATS Conf. pp. 4195–4205 (2020)

Data

- Publicly available real-world relational datasets
- Focus on a smaller subset of the original attributes

	#objs.	*sensitive* attribute	*non-sensitive* attributes
Adult	48 842	sex	age, fnlgwt, education_num, capital_gain, hours_per_week
Bank	40 004	marital	age, balance, duration
CreditCard	10 127	sex	customer_age, dependent_count, avg_utilization_ratio, total_relationship_count
Diabetes	101 763	sex	age, time_in_hospital
Student	649	sex	age, study_time, absences

Evaluation goals

$$inter(\mathcal{A}^{\neg F}) = \frac{1}{|\Theta|} \sum_{u,v \in \Theta} sim_{\mathcal{A}^{\neg F}}(u,v)$$

$$inter(\mathcal{A}^{F}) = \frac{1}{|\Theta|} \sum_{u,v \in \Theta} sim_{\mathcal{A}^{F}}(u,v)$$

$$intra(\mathcal{A}^{\neg F}) = \frac{1}{|\Omega|} \sum_{u,v \in \Omega} sim_{\mathcal{A}^{\neg F}}(u,v)$$

$$intra(\mathcal{A}^{F}) = \frac{1}{|\Omega|} \sum_{u,v \in \Omega} sim_{\mathcal{A}^{F}}(u,v)$$

$$\Omega = \{u, v \in \mathcal{X} | \mathcal{C}(u) = \mathcal{C}(v)\}, \Theta = \{u, v \in \mathcal{X} | \mathcal{C}(u) \neq \mathcal{C}(v)\}$$

Running times were measured while executing on the *Cresco6* cluster*

* https://www.eneagrid.enea.it

Hyper-params and Configurations

- Random sampling of the original data
  - 1k/10k tuples which preserve some desired ratio between the protected classes
- Specification of *p* and *q* parameters
  - *p/q* represents the minimum balance required by each cluster
- Minimum shared requirements, e.g., single and binary sensitive attribute
- Number of clusters *k* as the (rounded) avg. number of clusters returned by CCBounds in ten iterations

	p, q	split ratio	k_{avg}	k
Adult-1k	1,2	650/350	3.12	3
Bank-1k	1,2	650/350	3.48	3
Credit-Card-1k	1,6	800/200	5.6	6
Diabetes-1k	1,2	540/460	5.2	5
Student-1k	1,2	266/383	3.88	4
Adult-10k	1,2	6 500/3 500	2.96	3
Bank-10k	1,2	6 500/3 500	3.28	3
Credit-Card-10k	1,6	4 769/5 358	6.32	6
Diabetes-10k	1,2	5 400/4 600	6.44	6
Adult-Full	2,5	32 650/16 192	3.64	4
Bank-Full	2,5	12 790/27 214	3.64	4
Diabetes-Full	1,2	47 055/54 708	OOM	6

Results - Balance

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
Adult-1k	CCBounds	3.12	0.565	**0.007**	**0.685**	0.524	**0.415**	**0.334**	**< 1**
	FAIRLETS	3	0.805	**0.004**	0.585	**0.319**	0.596	**0.335**	**< 1**
	HST-FC	3	**0.971**	0.01	0.616	0.335	0.599	**0.336**	**< 1**
	SIGNED	41	0.66	0.03	0.59	**0.32**	0.60	**0.33**	240
Adult-10k	CCBounds	2.96	0.52	0.03	**0.65**	0.43	**0.43**	**0.33**	3.86
	FAIRLETS	3	0.82	**0.003**	0.60	**0.32**	0.615	**0.33**	**< 1**
	HST-FC	3	**0.98**	**0.006**	0.626	0.336	0.618	**0.336**	3.03
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Adult-Full	CCBounds	3.64	0.56	**0.003**	**0.69**	0.47	**0.42**	0.24	75.5
	FAIRLETS	4	0.66	0.02	0.59	**0.32**	0.62	**0.34**	**6.5**
	HST-FC	4	**0.96**	**0.008**	0.63	0.34	0.62	**0.34**	72.86
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-1k	CCBounds	3.48	0.565	**0.006**	**0.727**	0.587	**0.441**	**0.369**	**< 1**
	FAIRLETS	3	0.828	**0.002**	0.606	**0.354**	0.613	**0.364**	**< 1**
	HST-FC	3	**0.968**	**0.007**	0.621	0.365	0.617	**0.365**	**< 1**
	SIGNED	41	0.7	0.03	0.61	**0.35**	0.63	**0.36**	224
Bank-10k	CCBounds	3.28	0.52	**0.0007**	**0.78**	0.63	**0.45**	**0.36**	4.74
	FAIRLETS	3	0.7	0.001	0.59	**0.32**	0.63	**0.36**	**< 1**
	HST-FC	3	**0.969**	0.004	0.656	0.365	0.656	**0.365**	3.07
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-Full	CCBounds	3.64	0.55	**0.0004**	**0.72**	0.55	**0.45**	**0.37**	51.1
	FAIRLETS	4	0.68	0.001	0.62	**0.34**	0.65	0.36	**5.3**
	HST-FC	4	**0.94**	0.008	0.66	0.37	0.66	**0.37**	28
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h

- “Fairness-native” methods yield better balance scores
- CCBounds is aligned with its direct competing method in most cases
- On small yet heavily unbalanced datasets (i.e., CreditCard-1k with an 80:20 ratio), CCBounds achieves the second-best score, while other competing methods struggle
- Overall, the balance obtained by CCBounds in all evaluation scenarios ranges from 0.45 to 0.613

Results - Balance

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
CreditCard-1k	CCBounds	5.6	0.613	0.127	**0.6**	0.497	**0.46**	0.362	< 1
	FAIRLETS	6	0.4	0.042	0.485	**0.355**	0.486	**0.375**	< 1
	HST-FC	6	**0.756**	**0.026**	0.513	0.373	0.481	**0.377**	< 1
	SIGNED	171	0.56	0.1	0.56	0.41	0.49	**0.38**	173
CreditCard-10k	CCBounds	6.32	0.496	0.17	**0.6**	0.46	**0.46**	0.32	4.1
	FAIRLETS	6	0.94	**0.01**	0.497	**0.34**	0.49	**0.337**	< 1
	HST-FC	6	**0.955**	**0.013**	0.52	**0.337**	0.491	**0.337**	2.52
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-1k	CCBounds	5.2	0.45	0.33	**0.622**	0.519	**0.512**	0.352	< 1
	FAIRLETS	5	**0.92**	**0.015**	0.537	0.381	0.532	**0.385**	< 1
	HST-FC	5	0.872	0.05	0.585	0.386	0.529	**0.386**	< 1
	SIGNED	106	0.85	0.04	0.58	**0.36**	0.54	**0.38**	257
Diabetes-10k	CCBounds	6.44	0.48	0.22	**0.65**	0.54	**0.5**	0.36	4.72
	FAIRLETS	6	**0.92**	**0.01**	0.53	**0.38**	**0.53**	**0.39**	< 1
	HST-FC	6	0.799	0.065	0.59	**0.388**	**0.53**	**0.386**	2.84
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-Full	CCBounds	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	FAIRLETS	6	**0.93**	**0.01**	OOM	OOM	OOM	OOM	**22.2**
	HST-FC	6	0.81	0.06	OOM	OOM	OOM	OOM	761.2
	SIGNED	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Student-1k	CCBounds	3.88	0.51	0.10	**0.625**	0.463	**0.471**	0.224	< 1
	FAIRLETS	4	0.82	**0.013**	0.528	**0.339**	0.543	**0.357**	< 1
	HST-FC	4	**0.93**	0.024	0.563	0.357	0.541	**0.358**	< 1
	SIGNED	55	0.82	0.04	0.57	**0.34**	0.55	**0.36**	71

- “Fairness-native” methods yield better balance scores
- CCBounds is aligned with its direct competing method in most cases
- On small yet heavily unbalanced datasets (i.e., CreditCard-1k with an 80:20 ratio), CCBounds achieves the second-best score, while other competing methods struggle
- Overall, the balance obtained by CCBounds in all evaluation scenarios ranges from 0.45 to 0.613

Results - Average Euclidean Fairness

		#*clust.*	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
Adult-1k	CCBounds	3.12	0.565	**0.007**	**0.685**	0.524	**0.415**	**0.334**	**< 1**
	Fairlets	3	0.805	**0.004**	0.585	**0.319**	0.596	**0.335**	**< 1**
	HST-FC	3	**0.971**	0.01	0.616	0.335	0.599	**0.336**	**< 1**
	Signed	41	0.66	0.03	0.59	**0.32**	0.60	**0.33**	240
Adult-10k	CCBounds	2.96	0.52	0.03	**0.65**	0.43	**0.43**	**0.33**	3.86
	Fairlets	3	0.82	**0.003**	0.60	**0.32**	0.615	**0.33**	**< 1**
	HST-FC	3	**0.98**	**0.006**	0.626	0.336	0.618	**0.336**	3.03
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Adult-Full	CCBounds	3.64	0.56	**0.003**	**0.69**	0.47	**0.42**	0.24	75.5
	Fairlets	4	0.66	0.02	0.59	**0.32**	0.62	**0.34**	**6.5**
	HST-FC	4	**0.96**	**0.008**	0.63	0.34	0.62	**0.34**	72.86
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-1k	CCBounds	3.48	0.565	**0.006**	**0.727**	0.587	**0.441**	**0.369**	**< 1**
	Fairlets	3	0.828	**0.002**	0.606	**0.354**	0.613	**0.364**	**< 1**
	HST-FC	3	**0.968**	**0.007**	0.621	0.365	0.617	**0.365**	**< 1**
	Signed	41	0.7	0.03	0.61	**0.35**	0.63	**0.36**	224
Bank-10k	CCBounds	3.28	0.52	**0.0007**	**0.78**	0.63	**0.45**	**0.36**	4.74
	Fairlets	3	0.7	0.001	0.59	**0.32**	0.63	**0.36**	**< 1**
	HST-FC	3	**0.969**	0.004	0.656	0.365	0.656	**0.365**	3.07
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-Full	CCBounds	3.64	0.55	**0.0004**	**0.72**	0.55	**0.45**	**0.37**	51.1
	Fairlets	4	0.68	0.001	0.62	**0.34**	0.65	0.36	**5.3**
	HST-FC	4	**0.94**	0.008	0.66	0.37	0.66	**0.37**	28
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h

- CCBounds obtains very good scores under different scenarios
- Among the best-performing approaches for the *Adult-1k*, *Adult-Full* and *Bank-1k* datasets
- Outperforms all the other methods by an order of magnitude on *Bank-10k* and *Bank-Full*
- Performances worsen while considering the remaining datasets

Results - Average Euclidean Fairness

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
CreditCard-1k	CCBounds	5.6	0.613	0.127	**0.6**	0.497	**0.46**	0.362	**< 1**
	FAIRLETS	6	0.4	0.042	0.485	**0.355**	0.486	**0.375**	**< 1**
	HST-FC	6	**0.756**	**0.026**	0.513	0.373	0.481	**0.377**	**< 1**
	SIGNED	171	0.56	0.1	0.56	0.41	0.49	**0.38**	173
CreditCard-10k	CCBounds	6.32	0.496	0.17	**0.6**	0.46	**0.46**	0.32	4.1
	FAIRLETS	6	0.94	**0.01**	0.497	**0.34**	0.49	**0.337**	**< 1**
	HST-FC	6	**0.955**	**0.013**	0.52	**0.337**	0.491	**0.337**	2.52
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-1k	CCBounds	5.2	0.45	0.33	**0.622**	0.519	**0.512**	0.352	**< 1**
	FAIRLETS	5	**0.92**	**0.015**	0.537	0.381	0.532	**0.385**	**< 1**
	HST-FC	5	0.872	0.05	0.585	0.386	0.529	**0.386**	**< 1**
	SIGNED	106	0.85	0.04	0.58	**0.36**	0.54	**0.38**	257
Diabetes-10k	CCBounds	6.44	0.48	0.22	**0.65**	0.54	**0.5**	0.36	4.72
	FAIRLETS	6	**0.92**	**0.01**	0.53	**0.38**	**0.53**	**0.39**	**< 1**
	HST-FC	6	0.799	0.065	0.59	**0.388**	**0.53**	**0.386**	2.84
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-Full	CCBounds	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	FAIRLETS	6	**0.93**	**0.01**	OOM	OOM	OOM	OOM	**22.2**
	HST-FC	6	0.81	0.06	OOM	OOM	OOM	OOM	761.2
	SIGNED	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Student-1k	CCBounds	3.88	0.51	0.10	**0.625**	0.463	**0.471**	0.224	**< 1**
	FAIRLETS	4	0.82	**0.013**	0.528	**0.339**	0.543	**0.357**	**< 1**
	HST-FC	4	**0.93**	0.024	0.563	0.357	0.541	**0.358**	**< 1**
	SIGNED	55	0.82	0.04	0.57	**0.34**	0.55	**0.36**	71

- CCBounds obtains very good scores under different scenarios
- Among the best-performing approaches for the *Adult-1k*, *Adult-Full* and *Bank-1k* datasets
- Outperforms all the other methods by an order of magnitude on *Bank-10k* and *Bank-Full*
- Performances worsen while considering the remaining datasets

Results - Similarities

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
Adult-1k	CCBounds	3.12	0.565	**0.007**	**0.685**	0.524	**0.415**	**0.334**	**< 1**
	FAIRLETS	3	0.805	**0.004**	0.585	**0.319**	0.596	**0.335**	**< 1**
	HST-FC	3	**0.971**	0.01	0.616	0.335	0.599	**0.336**	**< 1**
	SIGNED	41	0.66	0.03	0.59	**0.32**	0.60	**0.33**	240
Adult-10k	CCBounds	2.96	0.52	0.03	**0.65**	0.43	**0.43**	**0.33**	3.86
	FAIRLETS	3	0.82	**0.003**	0.60	**0.32**	0.615	**0.33**	**< 1**
	HST-FC	3	**0.98**	**0.006**	0.626	0.336	0.618	**0.336**	3.03
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Adult-Full	CCBounds	3.64	0.56	**0.003**	**0.69**	0.47	**0.42**	0.24	75.5
	FAIRLETS	4	0.66	0.02	0.59	**0.32**	0.62	**0.34**	**6.5**
	HST-FC	4	**0.96**	**0.008**	0.63	0.34	0.62	**0.34**	72.86
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-1k	CCBounds	3.48	0.565	**0.006**	**0.727**	0.587	**0.441**	**0.369**	**< 1**
	FAIRLETS	3	0.828	**0.002**	0.606	**0.354**	0.613	**0.364**	**< 1**
	HST-FC	3	**0.968**	**0.007**	0.621	0.365	0.617	**0.365**	**< 1**
	SIGNED	41	0.7	0.03	0.61	**0.35**	0.63	**0.36**	224
Bank-10k	CCBounds	3.28	0.52	**0.0007**	**0.78**	0.63	**0.45**	**0.36**	4.74
	FAIRLETS	3	0.7	0.001	0.59	**0.32**	0.63	**0.36**	**< 1**
	HST-FC	3	**0.969**	0.004	0.656	0.365	0.656	**0.365**	3.07
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-Full	CCBounds	3.64	0.55	**0.0004**	**0.72**	0.55	**0.45**	**0.37**	51.1
	FAIRLETS	4	0.68	0.001	0.62	**0.34**	0.65	0.36	**5.3**
	HST-FC	4	**0.94**	0.008	0.66	0.37	0.66	**0.37**	28
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h

- On the sensitive attributes, CCBounds tends to group a few more objects with the same sensitive attribute value than the other methods
- CCBounds is still able to properly separate the objects into clusters, when accounting for the sensitive attribute
- CCBounds achieves the best performance in all the considered evaluation scenarios when considering non-sensitive attributes

Results - Similarities

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
CreditCard-1k	CCBounds	5.6	0.613	0.127	**0.6**	0.497	**0.46**	0.362	**< 1**
	Fairlets	6	0.4	0.042	0.485	**0.355**	0.486	**0.375**	**< 1**
	HST-FC	6	**0.756**	**0.026**	0.513	0.373	0.481	**0.377**	**< 1**
	Signed	171	0.56	0.1	0.56	0.41	0.49	**0.38**	173
CreditCard-10k	CCBounds	6.32	0.496	0.17	**0.6**	0.46	**0.46**	0.32	4.1
	Fairlets	6	0.94	**0.01**	0.497	**0.34**	0.49	**0.337**	**< 1**
	HST-FC	6	**0.955**	**0.013**	0.52	**0.337**	0.491	**0.337**	2.52
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-1k	CCBounds	5.2	0.45	0.33	**0.622**	0.519	**0.512**	0.352	**< 1**
	Fairlets	5	**0.92**	**0.015**	0.537	0.381	0.532	**0.385**	**< 1**
	HST-FC	5	0.872	0.05	0.585	0.386	0.529	**0.386**	**< 1**
	Signed	106	0.85	0.04	0.58	**0.36**	0.54	**0.38**	257
Diabetes-10k	CCBounds	6.44	0.48	0.22	**0.65**	0.54	**0.5**	0.36	4.72
	Fairlets	6	**0.92**	**0.01**	0.53	**0.38**	**0.53**	**0.39**	**< 1**
	HST-FC	6	0.799	0.065	0.59	**0.388**	**0.53**	**0.386**	2.84
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-Full	CCBounds	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Fairlets	6	**0.93**	**0.01**	OOM	OOM	OOM	OOM	**22.2**
	HST-FC	6	0.81	0.06	OOM	OOM	OOM	OOM	761.2
	Signed	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Student-1k	CCBounds	3.88	0.51	0.10	**0.625**	0.463	**0.471**	0.224	**< 1**
	Fairlets	4	0.82	**0.013**	0.528	**0.339**	0.543	**0.357**	**< 1**
	HST-FC	4	**0.93**	0.024	0.563	0.357	0.541	**0.358**	**< 1**
	Signed	55	0.82	0.04	0.57	**0.34**	0.55	**0.36**	71

- On the sensitive attributes, CCBounds tends to group a few more objects with the same sensitive attribute value than the other methods
- CCBounds is still able to properly separate the objects into clusters, when accounting for the sensitive attribute
- CCBounds achieves the best performance in all the considered evaluation scenarios when considering non-sensitive attributes

Results - Running Times

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
Adult-1k	CCBounds	3.12	0.565	**0.007**	**0.685**	0.524	**0.415**	**0.334**	< **1**
	FAIRLETS	3	0.805	**0.004**	0.585	**0.319**	0.596	**0.335**	< **1**
	HST-FC	3	**0.971**	0.01	0.616	0.335	0.599	**0.336**	< **1**
	SIGNED	41	0.66	0.03	0.59	**0.32**	0.60	**0.33**	240
Adult-10k	CCBounds	2.96	0.52	0.03	**0.65**	0.43	**0.43**	**0.33**	3.86
	FAIRLETS	3	0.82	**0.003**	0.60	**0.32**	0.615	**0.33**	< **1**
	HST-FC	3	**0.98**	**0.006**	0.626	0.336	0.618	**0.336**	3.03
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Adult-Full	CCBounds	3.64	0.56	**0.003**	**0.69**	0.47	**0.42**	0.24	75.5
	FAIRLETS	4	0.66	0.02	0.59	**0.32**	0.62	**0.34**	**6.5**
	HST-FC	4	**0.96**	**0.008**	0.63	0.34	0.62	**0.34**	72.86
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-1k	CCBounds	3.48	0.565	**0.006**	**0.727**	0.587	**0.441**	**0.369**	< **1**
	FAIRLETS	3	0.828	**0.002**	0.606	**0.354**	0.613	**0.364**	< **1**
	HST-FC	3	**0.968**	**0.007**	0.621	0.365	0.617	**0.365**	< **1**
	SIGNED	41	0.7	0.03	0.61	**0.35**	0.63	**0.36**	224
Bank-10k	CCBounds	3.28	0.52	**0.0007**	**0.78**	0.63	**0.45**	**0.36**	4.74
	FAIRLETS	3	0.7	0.001	0.59	**0.32**	0.63	**0.36**	< **1**
	HST-FC	3	**0.969**	0.004	0.656	0.365	0.656	**0.365**	3.07
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h
Bank-Full	CCBounds	3.64	0.55	**0.0004**	**0.72**	0.55	**0.45**	**0.37**	51.1
	FAIRLETS	4	0.68	0.001	0.62	**0.34**	0.65	0.36	**5.3**
	HST-FC	4	**0.94**	0.008	0.66	0.37	0.66	**0.37**	28
	SIGNED	NA	NA	NA	NA	NA	NA	NA	> 48h

- FAIRLETS, HST-FC and CCBounds guarantee reasonable running times
- CCBounds overcomes its direct competing method SIGNED
  - Parallelized pairwise similarity computation
  - Abnormal number of clusters for SIGNED

Results - Running Times

		#clust.	balance ↑	AE ↓	$intra(\mathcal{A}^{\neg F})$ ↑	$intra(\mathcal{A}^{F})$ ↓	$inter(\mathcal{A}^{\neg F})$ ↓	$inter(\mathcal{A}^{F})$ ↑	time (s) ↓
CreditCard-1k	CCBounds	5.6	0.613	0.127	**0.6**	0.497	**0.46**	0.362	< **1**
	Fairlets	6	0.4	0.042	0.485	**0.355**	0.486	**0.375**	< **1**
	HST-FC	6	**0.756**	**0.026**	0.513	0.373	0.481	**0.377**	< **1**
	Signed	171	0.56	0.1	0.56	0.41	0.49	**0.38**	173
CreditCard-10k	CCBounds	6.32	0.496	0.17	**0.6**	0.46	**0.46**	0.32	4.1
	Fairlets	6	0.94	**0.01**	0.497	**0.34**	0.49	**0.337**	< **1**
	HST-FC	6	**0.955**	**0.013**	0.52	**0.337**	0.491	**0.337**	2.52
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-1k	CCBounds	5.2	0.45	0.33	**0.622**	0.519	**0.512**	0.352	< **1**
	Fairlets	5	**0.92**	**0.015**	0.537	0.381	0.532	**0.385**	< **1**
	HST-FC	5	0.872	0.05	0.585	0.386	0.529	**0.386**	< **1**
	Signed	106	0.85	0.04	0.58	**0.36**	0.54	**0.38**	257
Diabetes-10k	CCBounds	6.44	0.48	0.22	**0.65**	0.54	**0.5**	0.36	4.72
	Fairlets	6	**0.92**	**0.01**	0.53	**0.38**	**0.53**	**0.39**	< **1**
	HST-FC	6	0.799	0.065	0.59	**0.388**	**0.53**	**0.386**	2.84
	Signed	NA	NA	NA	NA	NA	NA	NA	> 48h
Diabetes-Full	CCBounds	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
	Fairlets	6	**0.93**	**0.01**	OOM	OOM	OOM	OOM	**22.2**
	HST-FC	6	0.81	0.06	OOM	OOM	OOM	OOM	761.2
	Signed	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
Student-1k	CCBounds	3.88	0.51	0.10	**0.625**	0.463	**0.471**	0.224	< **1**
	Fairlets	4	0.82	**0.013**	0.528	**0.339**	0.543	**0.357**	< **1**
	HST-FC	4	**0.93**	0.024	0.563	0.357	0.541	**0.358**	< **1**
	Signed	55	0.82	0.04	0.57	**0.34**	0.55	**0.36**	71

- FAIRLETS, HST-FC and CCBounds guarantee reasonable running times
- CCBounds overcomes its direct competing method SIGNED
  - Parallelized pairwise similarity computation
  - Abnormal number of clusters for SIGNED

Conclusions

- We assessed how correlation clustering can handle fair clustering
- Experimental evidence that CCBounds may serve as a good trade-off between the traditional and fairness-aware clustering conditions

Future Work

- Alternative definitions of the similarity functions
- Generalization of CCBounds to
  - Multiple protected values
  - Multiple sensitive attributes with many values

Thanks!

Any questions?

You can find me at:

https://luciolcv.github.io/

lucio.lacava@dimes.unical.it

@luciolcw