https://iclr.cc/Conferences/2025/PosterInstructions

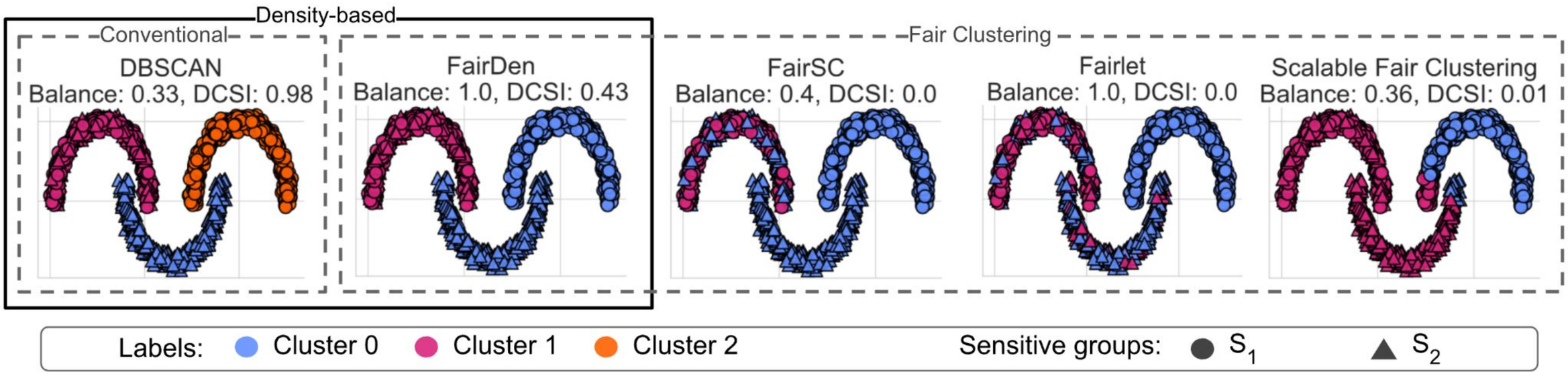
FairDen: Fair Density-Based Clustering

Lena Krieger*[1], Anna Beer*[2], Pernille Matthews[3], Anneka Myrup Thiesson[3], Ira Assent[1,3]

[1] IAS-8, Forschungszentrum Jülich [2] Faculty of Computer Science, University of Vienna [3] Department of Computer Science, Aarhus University

*Equal Contribution.

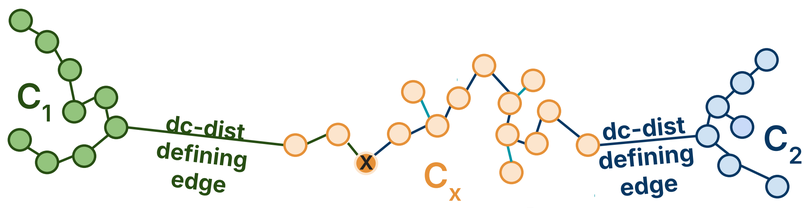
Fairness in Density-based Clustering



For a binary sensitive attribute indicated by the shape, a **balanced** density-based cluster has the **same ratio** of sensitive groups (circles and triangles) as in the overall dataset. The three equally sized moons are density-based clusters that have different ratios: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles. We show the balance and DCSI values (higher is better), for five clustering methods: FairDen achieves optimal balance and a high density-connectivity of clusters.

Method

Capture Density-connectivity structure:



The *density-connectivity distance* [1] is based on the mutual reachability distance *dm* that takes the density of points into account by using the points' core distances d_{core} :

Ensure balance of sensitive attributes:

$$f_p^{S_x} - (|S_x|/n) \cdot \mathbf{1}_n$$

a. Binary encoding for each *sensitive group*

	△	◯	◯	△
△	1	0	0	1
◯	0	1	1	0

b. Columns of fairness matrix

$$\triangle(1\ 0\ 0\ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

Combine density-connectivity and fairness:

$$\min_{\mathcal{H} \in \mathbb{R}^{n \times k}} Tr(\mathcal{H}^\top \mathcal{L} \mathcal{H})$$

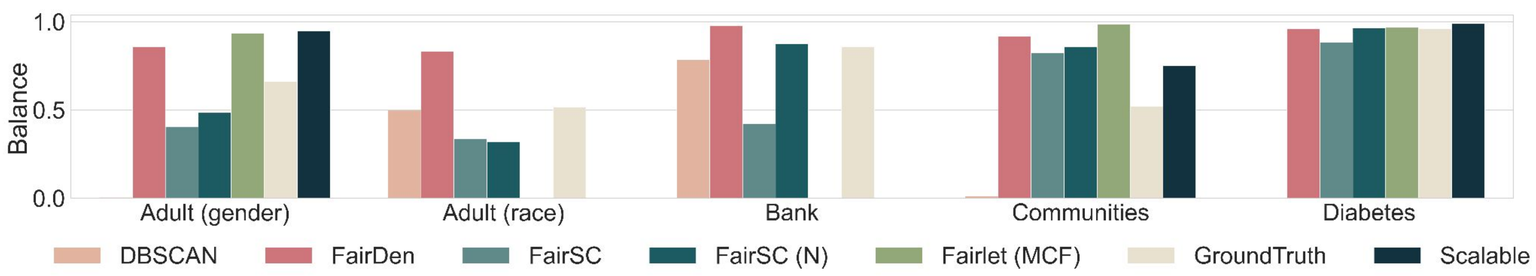
$$\text{subject to } \mathcal{H}^\top \mathcal{D} \mathcal{H} = \mathcal{I}_k \text{ and } \mathcal{F}^\top \mathcal{H} = 0$$

Imposing the fairness constraint ($\mathcal{F}^\top\mathcal{H}$) to a hierarchy of density-connected clusters transforms the problem into a ***graph-cut problem*** solvable with ***spectral clustering*** **[2]**.

Results

FairDen finds more balanced clusterings w.r.t. sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

Experiments regarding fairness: Balance values for all competitors on benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not always included as they cannot handle non-binary sensitive groups.



Experiments regarding clustering quality: Number of clusters ***k***, **Balance** (assessing group-level fairness), **DCSI** (assessing density-connectivity of clusters), and **ARI** (assessing similarity to ground truth clustering) for real-world benchmark data.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00

	k	Algorithm	Balance	DCSI	ARI
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

References

[1] Anna Beer, Andrew Draganov, Ellen Hohma, Philipp Jahn, Christian MM Frey, and Ira Assent. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 80-92. 2023.

[2] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. "Guarantees for spectral clustering with fairness constraints." In *International conference on machine learning*, pp. 3458-3467. PMLR, 2019.

[3] Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in neural information processing systems*, 30.

[4] Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019, May). Scalable fair clustering. In *International conference on machine learning* (pp. 405-413). PMLR.

Code

Paper

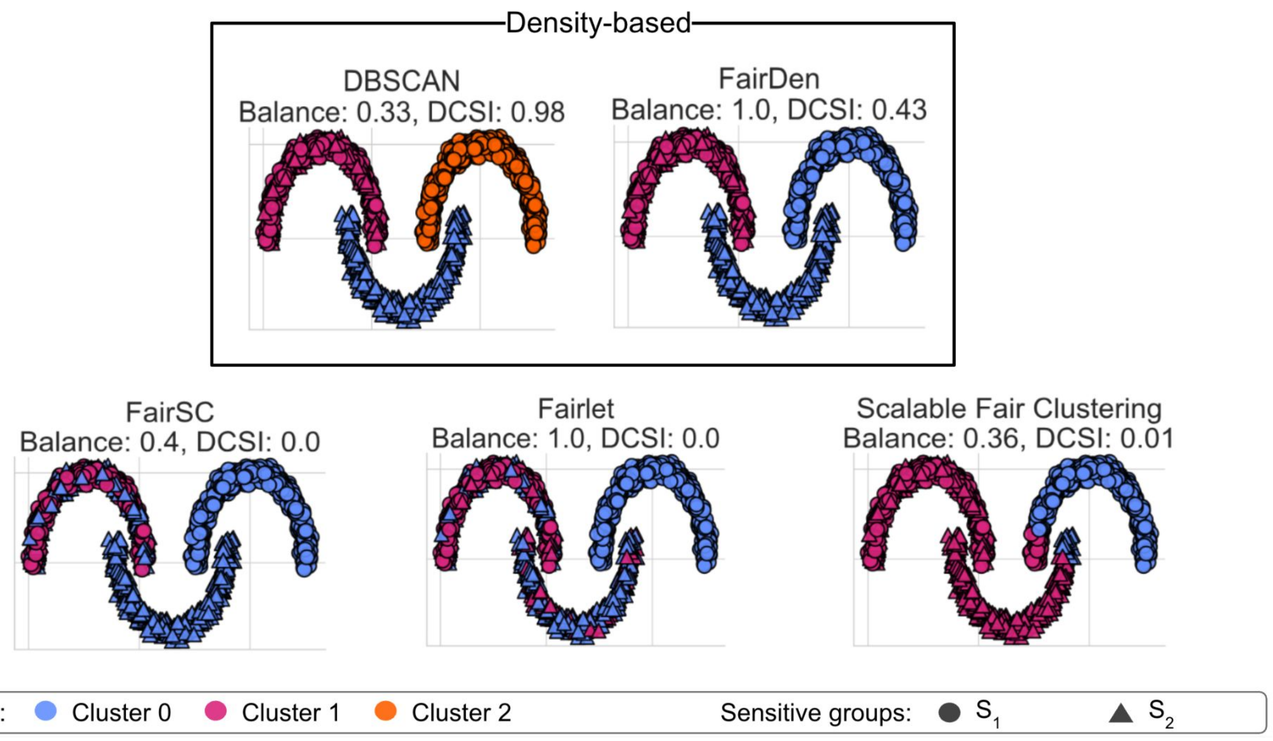
ECML PKDD 2025
Originally published at ICLR

FairDen: Fair Density-Based Clustering

Lena Krieger*[1], Anna Beer*[2], Pernille Matthews[3], Anneka Myrup Thiesson[3], Ira Assent[1,3]

[1] IAS-8, Forschungszentrum Jülich, Germany [2] Institute for Computer Science, LMU Munich, Germany
[3] Faculty of Computer Science, University of Vienna, Austria [4] Department of Computer Science, Aarhus University, Denmark
*Equal Contribution.

Fairness in Density-Based Clustering



For a binary sensitive attribute indicated by the shape, a **balanced** density-based cluster has the **same ratio** of sensitive groups (circles and triangles) as in the overall dataset. The three equally sized moons are density-based clusters that have different ratios: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles. We show the balance and DCSI values (higher is better), for five clustering methods: FairDen achieves optimal balance and a high density-connectivity of clusters.

Summary

FairDen is the first **density-based group-level fair** clustering algorithm.
It can even cluster data with:

- **multiple sensitive attributes**,
- **multiple sensitive groups**, and
- **categorical features.**

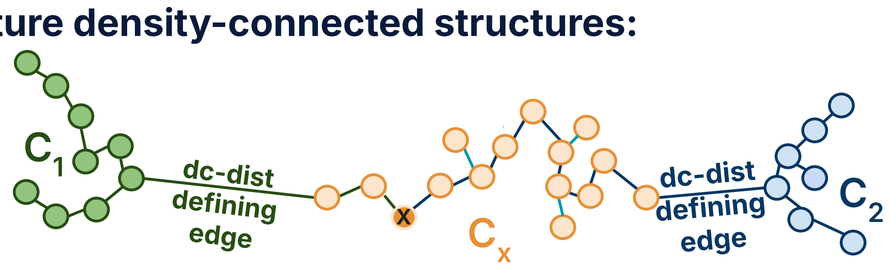
Related Work

Most other state-of-the-art group-level fair clustering methods cannot handle multiple sensitive attributes, multiple sensitive groups, or categorical values in the non-sensitive attributes.

Algorithm	Fairlet [3]	Scalable Fair Clustering [4]	Fair SC [2]	FairDen (Ours)
Density-based	✗	✗	✗	✓
Multiple sensitive attributes	✗	✗	✗	✓
Multiple (>2) sensitive groups	✗	✗	✓	✓
Categorical Features	✗	✗	✗	✓

Method

Capture density-connected structures:



The *density-connectivity distance* [1] is based on the mutual reachability distance d_m that takes the density of points into account by using the points' core distances d_{core}:

$$d_m(x,y) = \max\left(d_{core}(x), d_{core}(y), d_{eucl}(x,y)\right)$$

The longest edge e along the min-max path on the graph given by d_m indicates the density-connectivity between two points x and y

$$d_{dc}(x,y) = \min_{P\in\mathcal{P}} \max_{e\in p(x,y)} |e|$$

Ensure balance of sensitive attributes:

$$f_p^{S_x} - (|S_x|/n)\cdot \mathbf{1}_n$$

a. Binary encoding for each *sensitive group* S_x

	△	○	○	△
△	1	0	0	1
○	0	1	1	0

b. Columns of fairness matrix:

$$\triangle\ (1\ 0\ 0\ 1) - \frac{|\triangle|}{n}\cdot \mathbf{1}_n$$

Combine density-connectivity and fairness:

$$\min_{\mathcal{H}\in\mathbb{R}^{n\times k}} Tr(\mathcal{H}^\top \mathcal{L}\mathcal{H}) \text{ subject to}$$

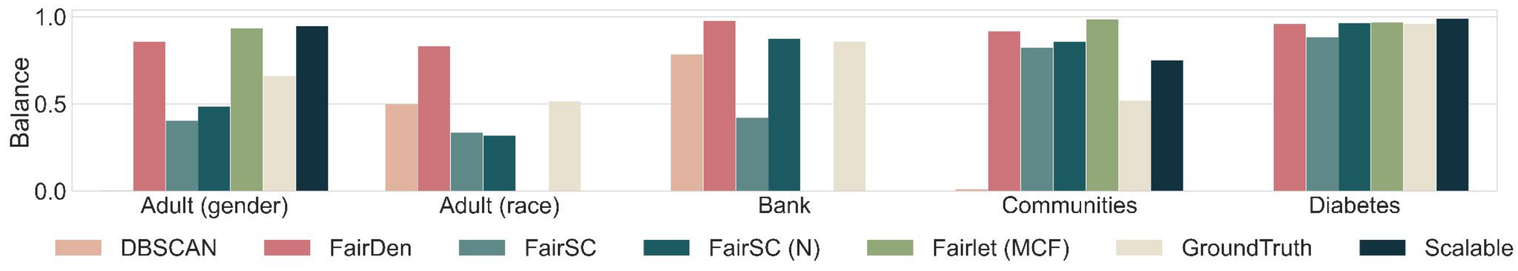
$$\mathcal{H}^\top \mathcal{D}\mathcal{H} = \mathcal{I}_k \text{ and } \mathcal{F}^\top\mathcal{H} = 0$$

Imposing the fairness constraint (F^TH) to a hierarchy of density-connected clusters transforms the problem into a ***graph-cut problem*** solvable with ***spectral clustering*** **[2]**.

Results

FairDen finds more balanced clusterings with respect to sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

Experiments regarding fairness: Balance values for all competitors on benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not always included as they cannot handle non-binary sensitive groups.



Experiments regarding clustering quality: Number of clusters **k**, **Balance** (assessing group-level fairness), **DCSI** (assessing density-connectivity of clusters), and **ARI** (assessing similarity to ground truth clustering) for real-world benchmark data.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00

	k	Algorithm	Balance	DCSI	ARI
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

References

[1] Anna Beer, Andrew Draganov, Ellen Hohma, Philipp Jahn, Christian MM Frey, and Ira Assent. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 80-92. 2023.
[2] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. "Guarantees for spectral clustering with fairness constraints." In *International conference on machine learning*, pp. 3458-3467. PMLR, 2019.
[3] Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in neural information processing systems*, 30.
[4] Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019, May). Scalable fair clustering. In *International conference on machine learning* (pp. 405-413). PMLR.

Code

Paper

FairDen: Fair Density-Based Clustering

Lena Krieger*[1], Anna Beer*[2], Pernille Matthews[3], Anneka Myrup Thiesson[3], Ira Assent[1,3]

[1] IAS-8, Forschungszentrum Jülich [2] Faculty of Computer Science, University of Vienna [3] Department of Computer Science, Aarhus University

*Equal Contribution.

Fairness in Density-based Clustering

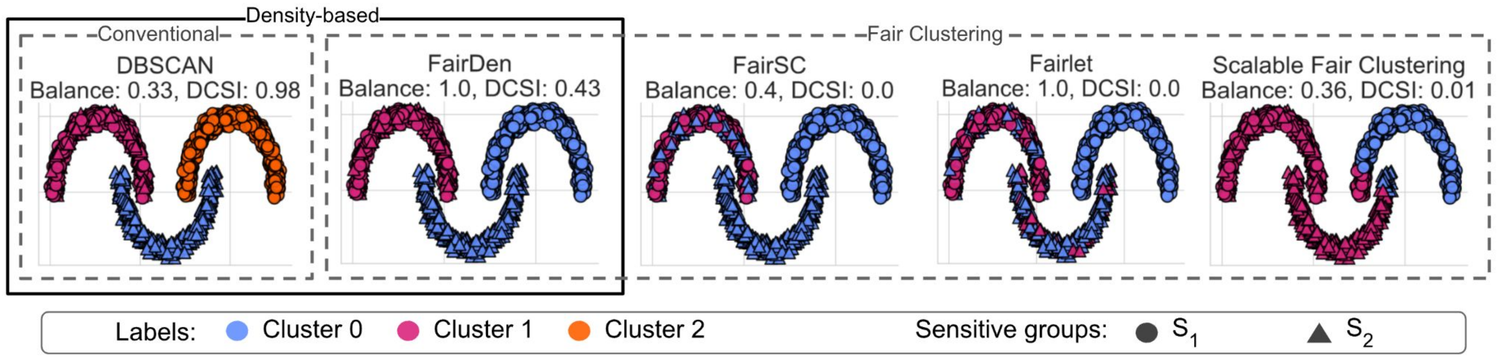


Figure 1: Top row: method, balance, and DCSI values (higher is better), respectively. A balanced density-based clustering has the same ratios of circles and triangles as in the overall dataset. Shapes and edge color indicate membership to one of two sensitive groups: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles, and all moons have the same number of data points.

Group-level Fair Clustering Approaches

Algorithm	Fairlets	Scalable Fair Clustering	Fair SC	FairDen (Ours)
Density-based	×	×	×	✓
Multiple sensitive attributes	×	×	×	✓
Multiple (>2) sensitive groups	×	×	✓	✓
Categorical Features	×	×	×	✓

Table 1. Properties of our group-level fair competitors. ***Sensitive attributes*** denote the name of the feature, e.g., gender, while ***sensitive groups*** encapsulate the values of the feature per group, e.g., female, male.

Method

Capture Density-connectivity structure:

$$d_m(x,y) = \max\left(d_{core}(x), d_{core}(y), d_{eucl}(x,y)\right)$$

$$d_{dc}(x,y) = \min_{P \in \mathcal{P}} \max_{e \in p(x,y)} |e|$$

a. Binary encoding for each *sensitive group*

△ ○ ○ △
△ (1 0 0 1)
○ (0 1 1 0)

b. Columns of fairness matrix

$$\triangle(1\ 0\ 0\ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

$$\min_{\mathcal{H} \in \mathbb{R}^{n \times k}} Tr(\mathcal{H}^\top \mathcal{L} \mathcal{H})$$

$$\text{subject to } \mathcal{H}^\top \mathcal{D} \mathcal{H} = \mathcal{I}_k \text{ and } \mathcal{F}^\top \mathcal{H} = 0$$

Results

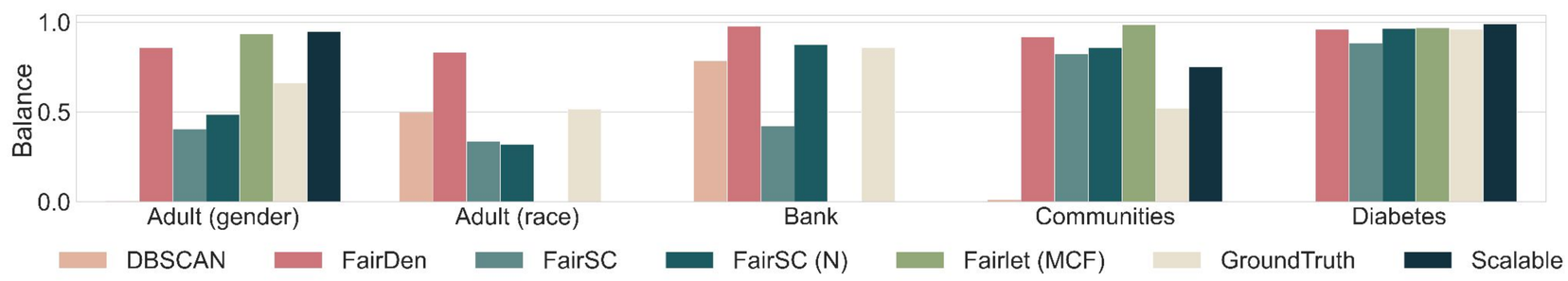


Figure 3: Balances for all competitors and benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not included for settings including non-binary sensitive groups.

FairDen determines more balanced clusterings w.r.t. sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

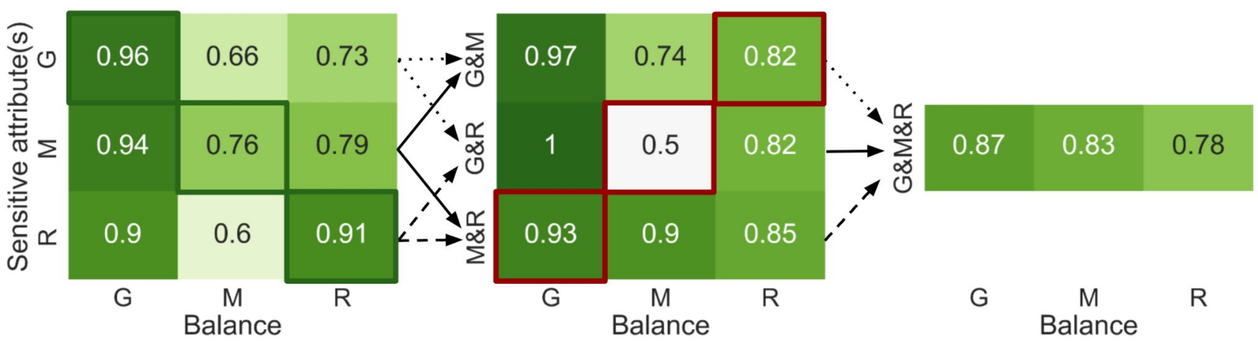


Figure 4: Columns show balance with respect to sensitive attributes *gender* (G), *marital status* (M) and *race* (R) in FairDen clusterings of the Adult dataset. **Left**: results for clustering when including only one of the sensitive attributes, *highest* column-wise values/balance for the attribute labeled as sensitive is framed in green. **Middle**: results for clusterings when including two of the sensitive attributes as indicated by the row description. *Lowest* column-wise values/ balance of attribute that is *not* labeled as sensitive is framed in red. **Right**: results for including all three sensitive attributes.

In contrast to state-of-the-art competitors, FairDen inherently handles categorical attributes, noise, and data with several sensitive attributes or groups.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

Table 2: Number of clusters k, Balance, DCSI, ARI for real-world benchmark data. Diabetes dataset for k=2 (Ground truth) and k=4 (DBSCAN clusters).

References

[1] Beer, Anna, et al. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.

Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." International Conference on Machine Learning. PMLR, 2019.

[2]

Code **Paper**

FairDen: Fair Density-Based Clustering

Lena Krieger*[1], Anna Beer*[2], Pernille Matthews[3], Anneka Myrup Thiesson[3], Ira Assent[1,3]

[1]IAS-8, Forschungszentrum Jülich [2]Faculty of Computer Science, University of Vienna [3]Department of Computer Science, Aarhus University

Fairness in Density-based Clustering

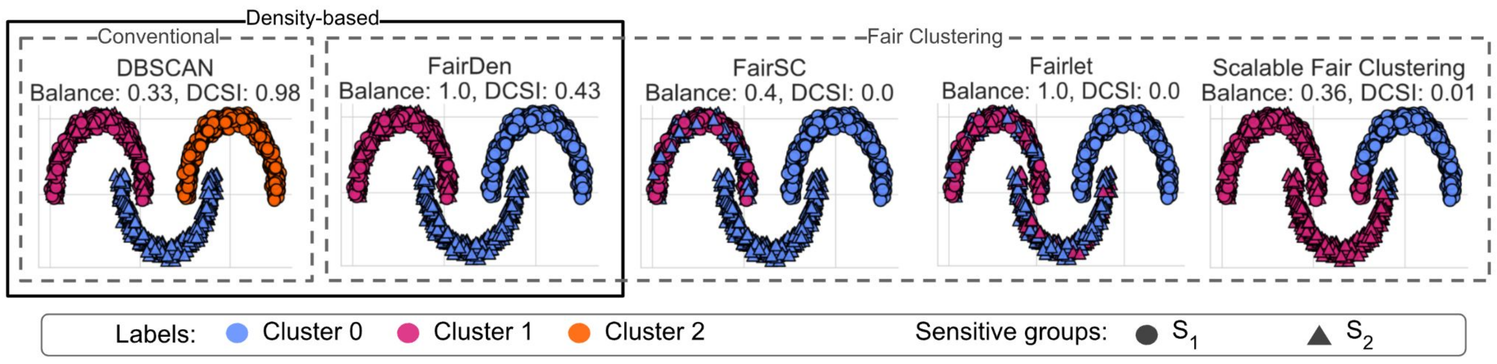


Figure 1: Top row: method, balance, and DCSI values (higher is better), respectively. A balanced density-based clustering has the same ratios of circles and triangles as in the overall dataset. Shapes and edge color indicate membership to one of two sensitive groups: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles, and all moons have the same number of data points.

Group-level Fair Clustering Approaches

Algorithm	Fairlets	Scalable Fair Clustering	Fair SC	FairDen (Ours)
Density-based	×	×	×	✓
Multiple sensitive attributes	×	×	×	✓
Multiple (>2) sensitive groups	×	×	✓	✓
Categorical Features	×	×	×	✓

Table 1. Properties of our group-level fair competitors. ***Sensitive attributes*** denote the name of the feature, e.g., gender, while ***sensitive groups*** encapsulate the values of the feature per group, e.g., female, male.

Method

1) Capture the density-connectivity structure

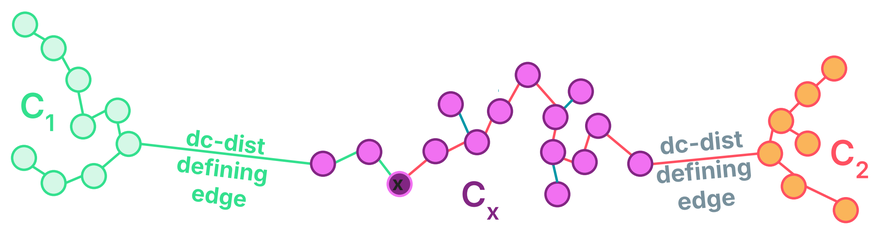


Figure 2: Regarding the *Density-connectivity distance* **[1]** C_2 is closer to x than C_1.

2) Ensure balance with fairness matrix

a. Binary encoding for each *sensitive group*

△ (1 0 0 1)
● (0 1 1 0)

b. Columns of fairness matrix

$$\triangle\,(1\ 0\ 0\ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

3) Combine both aspects

Imposing the fairness constraint to a hierarchy of density-connected clusters transforms the problem into a ***graph-cut problem*** solvable with ***spectral clustering*** **[2]**.

Results

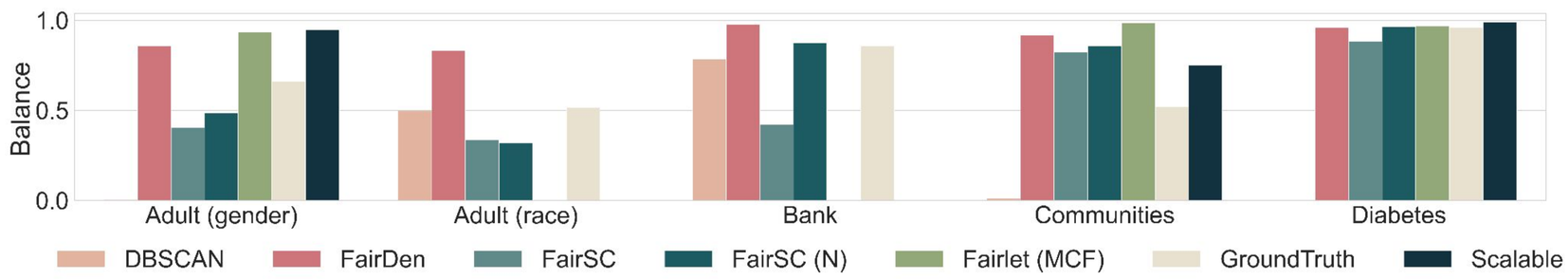


Figure 3: Balances for all competitors and benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not included for settings including non-binary sensitive groups.

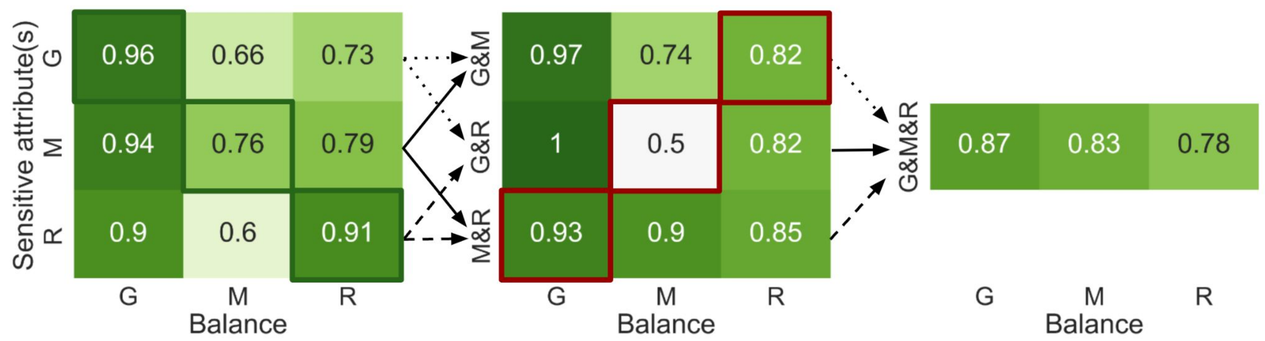


Figure 4: Columns show balance with respect to sensitive attributes *gender* (G), *marital status* (M) and *race* (R) in FairDen clusterings of the Adult dataset. **Left**: results for clustering when including only one of the sensitive attributes, *highest* column-wise values/balance for the attribute labeled as sensitive is framed in green. **Middle**: results for clusterings when including two of the sensitive attributes as indicated by the row description. *Lowest* column-wise values/ balance of attribute that is *not* labeled as sensitive is framed in red. **Right**: results for including all three sensitive attributes.

FairDen determines more balanced clusterings w.r.t. sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

In contrast to state-of-the-art competitors, FairDen inherently handles categorical attributes, noise, and data with several sensitive attributes or groups.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

Table 2: Number of clusters k, Balance, DCSI, ARI for real-world benchmark data. Diabetes dataset for k=2 (Ground truth) and k=4 (DBSCAN clusters).

References

[1] Beer, Anna, et al. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.

[2] Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." International Conference on Machine Learning. PMLR, 2019.

Code

Paper

FairDen: Fair Density-Based Clustering

Lena Krieger*[1], Anna Beer*[2], Pernille Matthews[3], Anneka Myrup Thiesson[3], Ira Assent[1,3]

[1]Forschungszentrum Jülich [2]University of Vienna [3]Aarhus University

Fairness in Density-based Clustering

Density-based Clustering (DBSCAN)

1. Determine core points
2. Select random (core) point to start a new cluster C_i
3. Merge all points with distance < ε from C_i to C_i
4. Assign non-core points:
   a. To C_i if distance to C_i < ε
   b. Noise otherwise

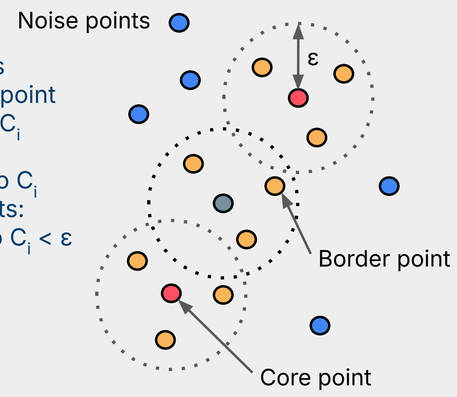


Figure 1: Illustration of DBSCAN workflow when *minPts=4*.

Group-level Fair / Balanced Clustering

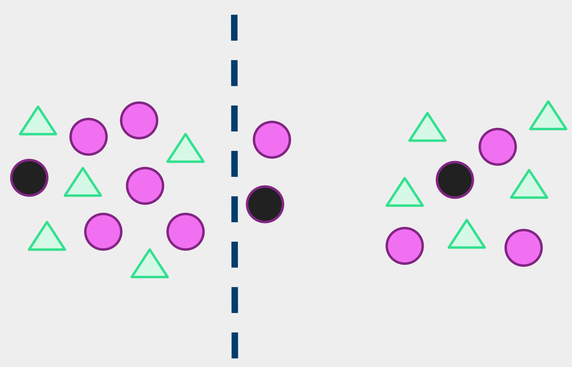


Figure 2: Line illustrates cut for group-level fair clustering. ***Sensitive groups*** are denoted in green and pink. This dataset includes only one ***sensitive attribute*** with two ***sensitive groups***.

Group-level Fair Clustering Approaches

Algorithm	Fairlets	Scalable Fair Clustering	Fair SC	FairDEN (Ours)
Density-based	×	×	×	✓
Multiple sensitive attributes	×	×	×	✓
Multiple (>2) sensitive groups	×	×	✓	✓
Categorical Features	×	×	×	✓

Table 1. Properties of our group-level fair competitors. ***Sensitive attributes*** denote the name of the feature, e.g., gender, while ***sensitive groups*** encapsulate the values of the feature per group, e.g., female, male.

Method

1) Capture the density-connectivity structure

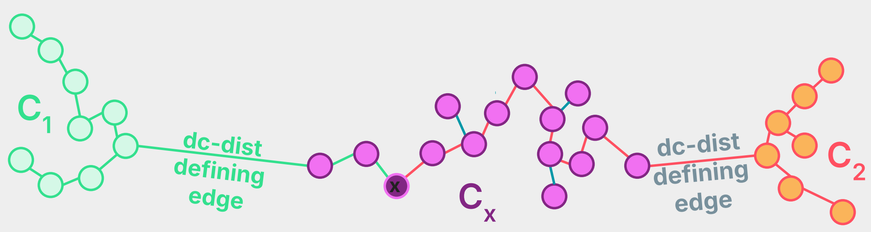


Figure 3: Regarding the *Density-connectivity distance* **[1]** C_2 is closer to x than C_1.

2) Ensure balance with fairness matrix

a. Binary encoding for each *sensitive group*

△ (1 0 0 1)
● (0 1 1 0)

b. Columns of fairness matrix

$$\triangle\ (1\ 0\ 0\ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

3) Combine both aspects

Imposing the fairness constraint to a hierarchy of density-connected clusters transforms the problem into a ***graph-cut problem*** solvable with ***spectral clustering*** **[2]**.

Results

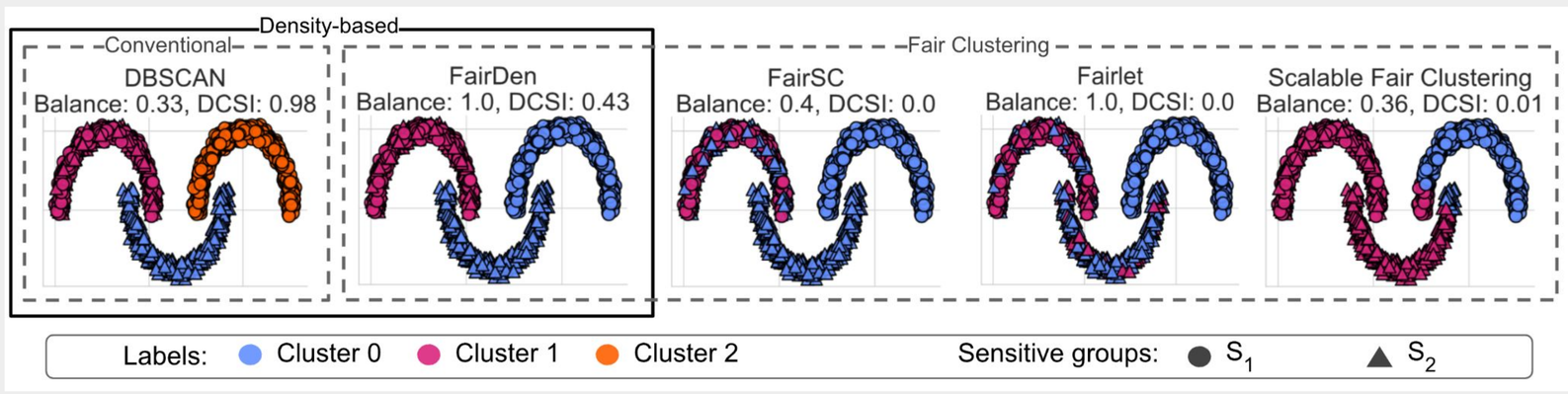


Figure 4: Top row: method, balance, and DCSI values (higher is better), respectively. A balanced density-based clustering has the same ratios of circles and triangles as in the overall dataset. Shapes and edge color indicate membership to one of two sensitive groups: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles, and all moons have the same number of data points.

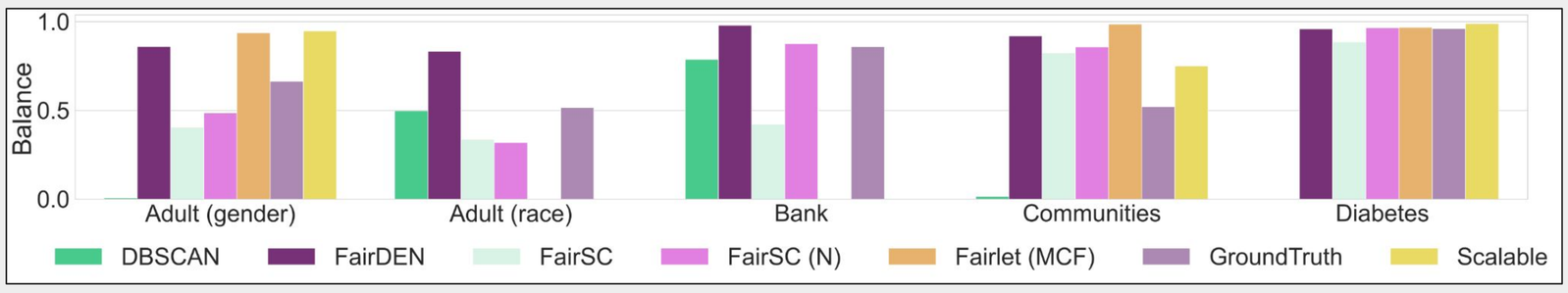


Figure 5: Balances for all competitors and benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not included for settings including non-binary sensitive groups.

FairDEN determines more balanced clusterings w.r.t. sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

In contrast to state-of-the-art competitors, FairDEN inherently handles categorical attributes, noise, and data with several sensitive attributes or groups.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

Table 2: Number of clusters k, Balance, DCSI, ARI for real-world benchmark data. Diabetes dataset for k=2 (Ground truth) and k=4 (DBSCAN clusters).

References

[1] Beer, Anna, et al. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.

[2] Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." International Conference on Machine Learning. PMLR, 2019.

Code:

Paper:

FairDEN: Fair Density-Based Clustering

Lena Krieger*, Anna Beer*, Pernille Matthews, Anneka Myrup Thiesson, Ira Assent

Fairness in Density-based Clustering

Density-based Clustering (DBSCAN)

1. Determine core points
2. Select random (core) point to start a new cluster C_i
3. Merge all points with distance < ε from C_i to C_i
4. Assign non-core points:
   a. To C_i if distance to C_i < ε
   b. Noise otherwise

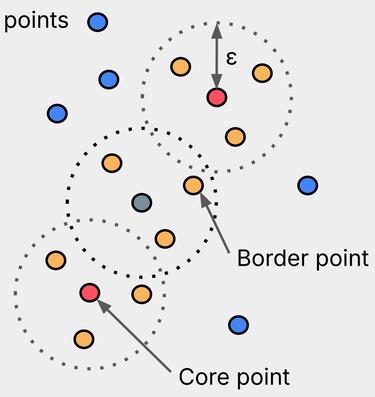


Figure 1: Illustration of DBSCAN workflow when *minPts=4*.

Group-level Fair / Balanced Clustering

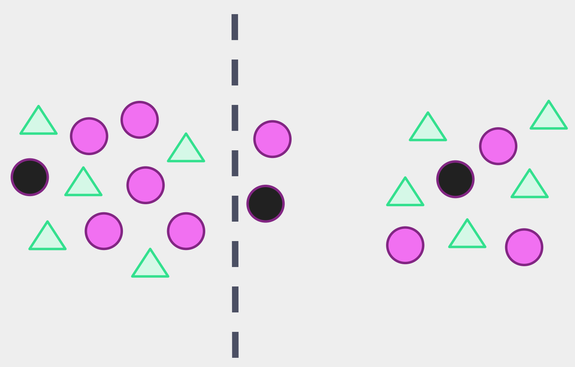


Figure 2: Line illustrates cut for group-level fair clustering. ***Sensitive groups*** are denoted in green and pink. This dataset includes only one ***sensitive attribute*** with two ***sensitive groups***.

Group-level Fair Clustering Approaches

Algorithm	Fairlets	Scalable Fair Clustering	Fair SC	FairDEN (Ours)
Density-based	✗	✗	✗	✓
Multiple sensitive attributes	✗	✗	✗	✓
Multiple (>2) sensitive groups	✗	✗	✓	✓
Categorical Features	✗	✗	✗	✓

Table 1. Properties of our group-level fair competitors. ***Sensitive attributes*** denote the name of the feature, e.g., gender, while ***sensitive groups*** encapsulate the values of the feature per group, e.g., female, male.

Method

1) Capture the density-connectivity structure

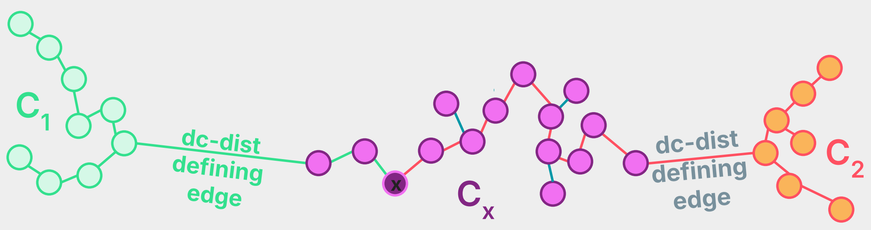


Figure 3: Regarding the *Density-connectivity distance* **[1]** C_2 is closer to x than C_1.

2) Ensure balance with fairness matrix

a. Binary encoding for each *sensitive group*

△ (1 0 0 1)
● (0 1 1 0)

b. Columns of fairness matrix

$$\triangle\,(1\ 0\ 0\ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

3) Combine both aspects

Imposing the fairness constraint to a hierarchy of density-connected clusters transforms the problem into a ***graph-cut problem*** solvable with ***spectral clustering*** **[2]**.

Results

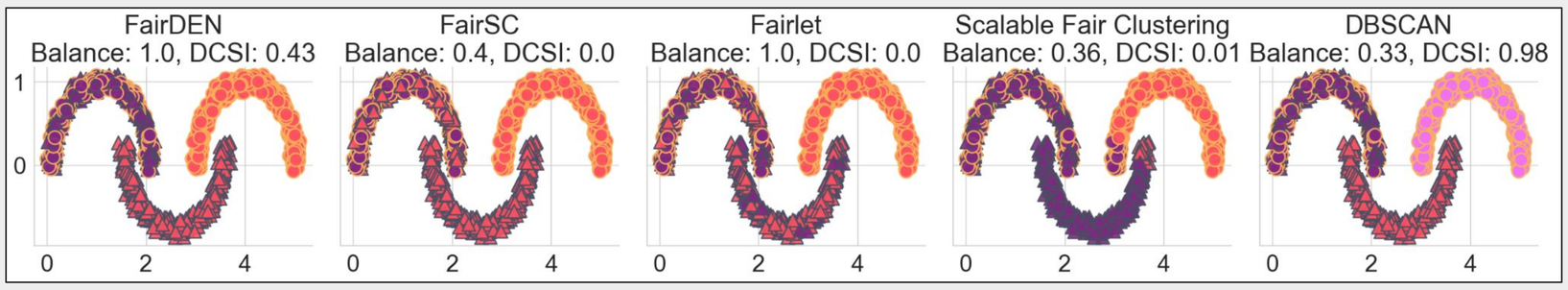


Figure 4: Top row: method, balance, and DCSI values (higher is better), respectively. A balanced density-based clustering has the same ratios of circles and triangles as in the overall dataset. Shapes and edge color indicate membership to one of two sensitive groups: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles, and all moons have the same number of data points.

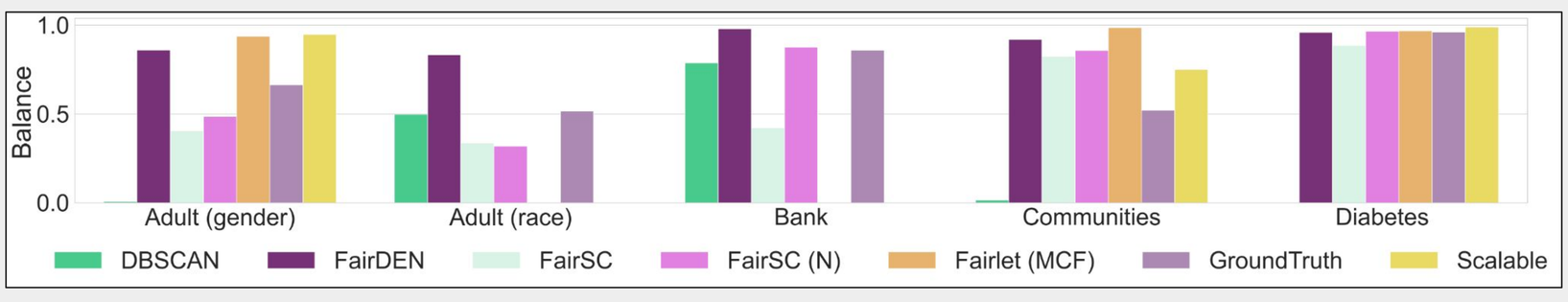


Figure 5: Balances for all competitors and benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not included for settings including non-binary sensitive groups.

FairDEN determines more balanced clusterings w.r.t. sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

In contrast to state-of-the-art competitors, FairDEN inherently handles categorical attributes, noise, and data with several sensitive attributes or groups.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

Table 2: Number of clusters k, Balance, DCSI, ARI for real-world benchmark data. Diabetes dataset for k=2 (Ground truth) and k=4 (DBSCAN clusters).

References

[1] Beer, Anna, et al. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.
[2] Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." International Conference on Machine Learning. PMLR, 2019.

AARHUS UNIVERSITY
JÜLICH Forschungszentrum
universität wien

Density-Based Clustering with fairness constraints

Lena Krieger*, Anna Beer*, Pernille Matthews, Anneka Myrup Thiesson, Ira Assent

Fairness in Density-based Clustering

Density-based Clustering (DBSCAN)

1. Determine core points
2. Select random (core) point to start a new cluster C_i
3. Merge all points with distance < ε from C_i to C_i
4. Assign non-core points:
   a. To C_i if distance to C_i < ε
   b. Noise otherwise

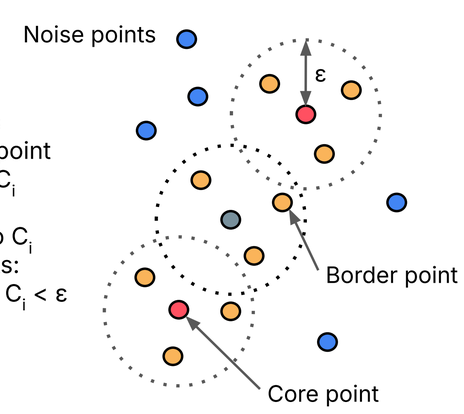


Figure 1: Illustration of DBSCAN workflow when *minPts=3*.

Group-level Fair / Balanced Clustering

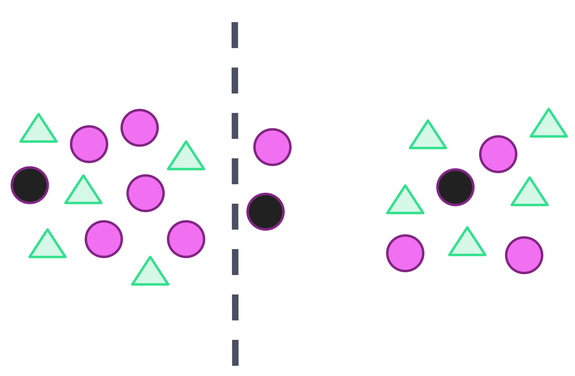


Figure 2: Line illustrates cut for group-level fair clustering. ***Sensitive groups*** are denoted in green and pink. This dataset includes only one ***sensitive attribute*** with two ***sensitive groups***.

Group-level Fair Clustering Approaches

Algorithm	Fairlets	Scalable Fair Clustering	Fair SC	FairDEN (Ours)
Density-based	×	×	×	✓
Multiple sensitive attributes	×	×	×	✓
Multiple (>2) sensitive groups	×	×	✓	✓
Categorical Features	×	×	×	✓

Table 1. Properties of our group-level fair competitors. ***Sensitive attributes*** denote the name of the feature, e.g., gender, while ***sensitive groups*** encapsulate the values of the feature per group, e.g., female, male.

Method

1) Capture the density-connectivity structure

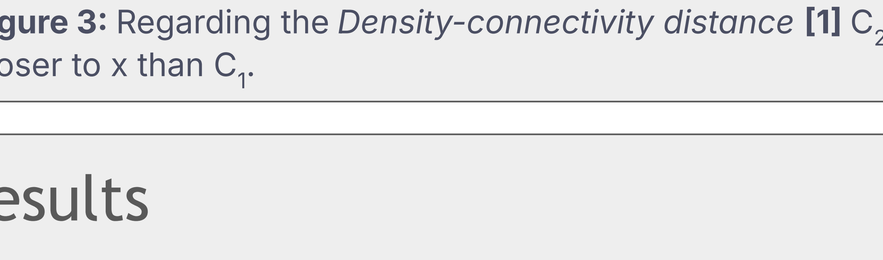


Figure 3: Regarding the *Density-connectivity distance* **[1]** C_2 is closer to x than C_1.

2) Ensure balance with fairness matrix

a. Binary encoding for each *sensitive group*

△ (1 0 0 1)
● (0 1 1 0)

b. Columns of fairness matrix

$$\triangle\,(1\ 0\ 0\ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

3) Combine both aspects

Imposing the fairness constraint to a hierarchy of density-connected clusters transforms the problem into a ***graph-cut problem*** solvable with ***spectral clustering*** **[2]**.

Results

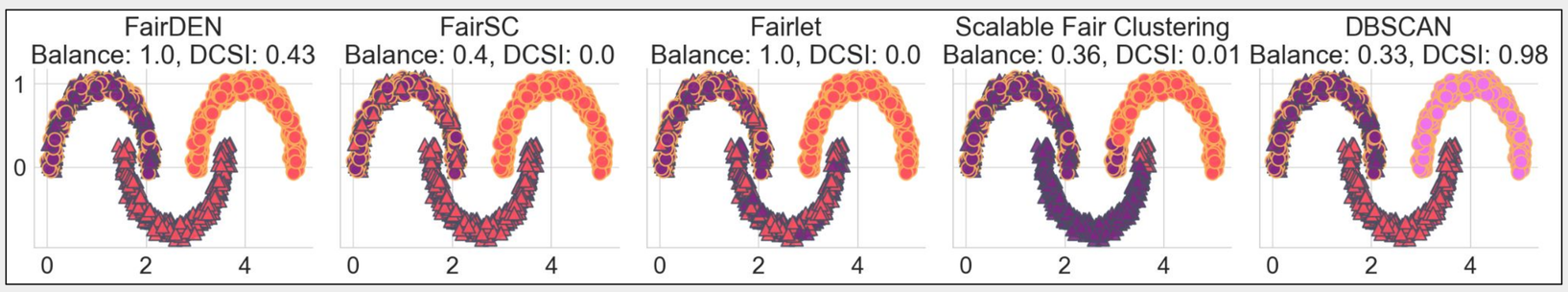


Figure 4: Top row: method, balance, and DCSI values (higher is better), respectively. A balanced density-based clustering has the same ratios of circles and triangles as in the overall dataset. Shapes and edge color indicate membership to one of two sensitive groups: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles, and all moons have the same number of data points.

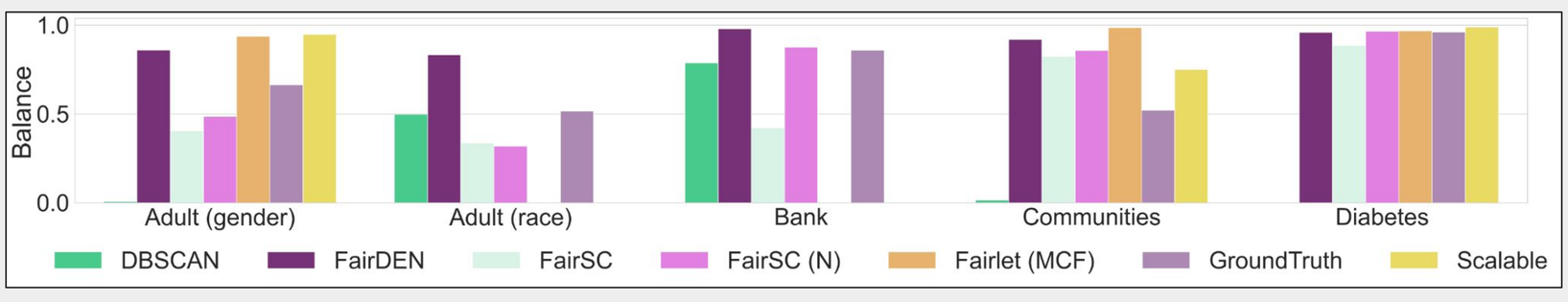


Figure 5: Balances for all competitors and benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not included for settings including non-binary sensitive groups.

FairDEN determines more balanced clusterings w.r.t. sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

In contrast to state-of-the-art competitors, FairDEN inherently handles categorical attributes, noise, and data with several sensitive attributes or groups.

	k	Algorithm	Balance	DCSI	ARI
Adult (gender)	2	DBSCAN	0.01	**0.97**	0.00
	2	FairDen	0.86	0.04	0.05
	2	FairSC	0.40	0.00	0.23
	2	FairSC (N)	0.49	0.00	0.27
	2	Fairlet (MCF)	0.94	0.00	0.00
	2	GroundTruth	0.66	0.00	1.00
	2	Scalable	**0.95**	0.01	-0.01
Adult (race)	2	DBSCAN	0.50	**0.99**	0.02
	2	FairDen	**0.83**	0.09	0.05
	2	FairSC	0.34	0.00	-0.03
	2	FairSC (N)	0.32	0.00	0.16
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.52	0.00	1.00
Bank	2	DBSCAN	0.79	**0.99**	0.01
	2	FairDen	**0.98**	0.14	0.21
	2	FairSC	0.42	0.00	-0.06
	2	FairSC (N)	0.88	0.00	-0.04
	2	Fairlet (MCF)	-	-	-
	2	Scalable	-	-	-
	2	GroundTruth	0.86	0.00	1.00
Communities	2	DBSCAN	0.01	**0.65**	-0.03
	2	FairDen	0.92	0.15	0.09
	2	FairSC	0.82	0.13	0.03
	2	FairSC (N)	0.86	0.13	0.03
	2	Fairlet (MCF)	**0.99**	0.05	0.16
	2	Scalable	0.75	0.08	-0.03
	2	GroundTruth	0.52	0.07	1.00
Diabetes	2	DBSCAN	-	-	-
	2	FairDen	0.96	**0.08**	0.01
	2	FairSC	0.89	0.00	-0.01
	2	FairSC (N)	0.96	0.00	0.01
	2	Fairlet (MCF)	0.97	0.00	0.00
	2	Scalable	**0.99**	0.01	0.00
	2	GroundTruth	0.96	0.00	1.00
	4	DBSCAN	0.01	**0.88**	-
	4	FairDen	0.95	0.24	-
	4	FairSC	0.23	0.01	-
	4	FairSC (N)	0.61	0.19	-
	4	Fairlet (MCF)	0.95	0.00	-
	4	Scalable	**0.96**	0.07	-
	4	GroundTruth	-	-	-

Table 2: Number of clusters k, Balance, DCSI, ARI for real-world benchmark data. Diabetes dataset for k=2 (Ground truth) and k=4 (DBSCAN clusters).

References

[1] Beer, Anna, et al. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023.
[2] Kleindessner, Matthäus, et al. "Guarantees for spectral clustering with fairness constraints." International Conference on Machine Learning. PMLR, 2019.

AARHUS UNIVERSITY
JÜLICH Forschungszentrum
universität wien