Reproducibilty Study of "Robust Fair Clustering: A Novel Fairness Attack and Defense Framework''

Authors: Iason Skylitsis, Zheng Feng, Idries Nasim, Camille Niessink

Supervisor: Luca Pantea

Overview

Context

Scope of Reproducibility

Methodology

Results

Discussion

Limitations and Future Work

Context

Clustering Algorithms

- Play an important role in the analysis and interpretation of vast amount of data
- Used in a variety of societal applications
  - This highlights critical issue of fairness
  - Fair Clustering Algorithms

Fair Clustering Algorithms

- Have not been explored from an adversarial attack perspective
  - Adversarial attacks aim to compromise utility of fairness

Fair Clustering Algorithms

- Have not been explored from an adversarial attack perspective
  - Adversarial attacks aim to compromise utility of fairness

- Experiment with adversarial attack to assess vulnerability of fair clustering algorithms
- Propose novel model designed to be highly resilient to the proposed fairness attack

Scope of Reproducibility

Claim 1

The *black-box* adversarial attack outlined in the original paper is capable of degrading the fairness performance, by perturbing a small percentage of protected group memberships, in the examined fair clustering models: Fair K-Center, Fair Spectral Clustering, and Scalable Fairlet Decomposition.

Claim 2

Fair K-Center, Fair Spectral Clustering, and Scalable Fairlet Decomposition, demonstrate a lack of robustness to adversarial influence, exhibiting significant volatility in terms of fairness utility metrics such as Balance and Entropy.

Claim 3

Consensus Fair Clustering exhibits high resilience against the proposed fairness attack, offering a robust solution for achieving fair clustering.

Methodology

Model Description

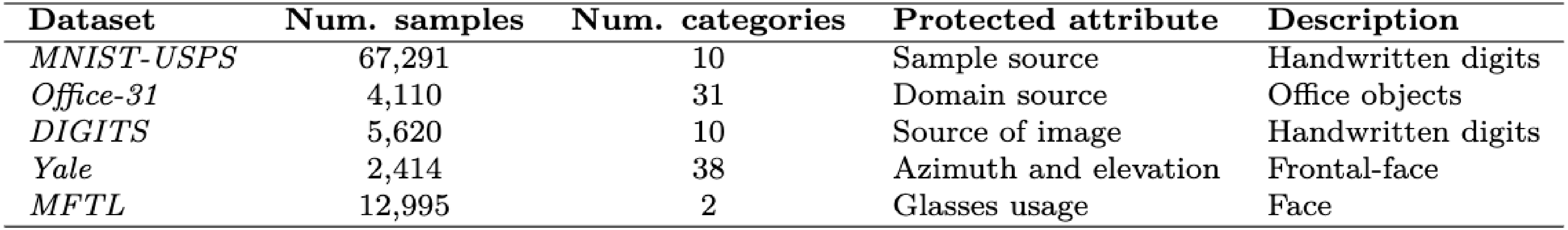
Three state-of-the-art fair clustering algorithms:

- Fair K-Center (KFC)
- Fair Spectral Clustering (FSC)
- Scalable Fairlet Decomposition (SFD)

Robust fair clustering algorithm (introduced by authors):

- Consensus Fair Clustering (CFC)

Datasets



Dataset	Num. samples	Num. categories	Protected attribute	Description
MNIST-USPS	67,291	10	Sample source	Handwritten digits
Office-31	4,110	31	Domain source	Office objects
DIGITS	5,620	10	Source of image	Handwritten digits
Yale	2,414	38	Azimuth and elevation	Frontal-face
MFTL	12,995	2	Glasses usage	Face

Table 1: Description of the datasets used in our experimentation.

Experimental results of reproducibilty study

Reproduction Results

Claim 1 The *black-box* adversarial attack outlined in the original paper is capable of degrading the fairness performance, by perturbing a small percentage of protected group memberships, in the examined fair clustering models: Fair K-Center, Fair Spectral Clustering, and Scalable Fairlet Decomposition.

Claim 2: Fair K-Center, Fair Spectral Clustering, and Scalable Fairlet Decomposition, demonstrate a lack of robustness to adversarial influence, exhibiting significant volatility in terms of fairness utility metrics such as Balance and Entropy.

Claim 3: *Consensus Fair Clustering* exhibits high resilience against the proposed fairness attack, offering a robust solution for achieving fair clustering.

Reproduction Results

Claim 1 The *black-box* adversarial attack outlined in the original paper is capable of degrading the fairness performance, by perturbing a small percentage of protected group memberships, in the examined fair clustering models: Fair K-Center, Fair Spectral Clustering, and Scalable Fairlet Decomposition.

→ PARTIALLY REPRODUCED

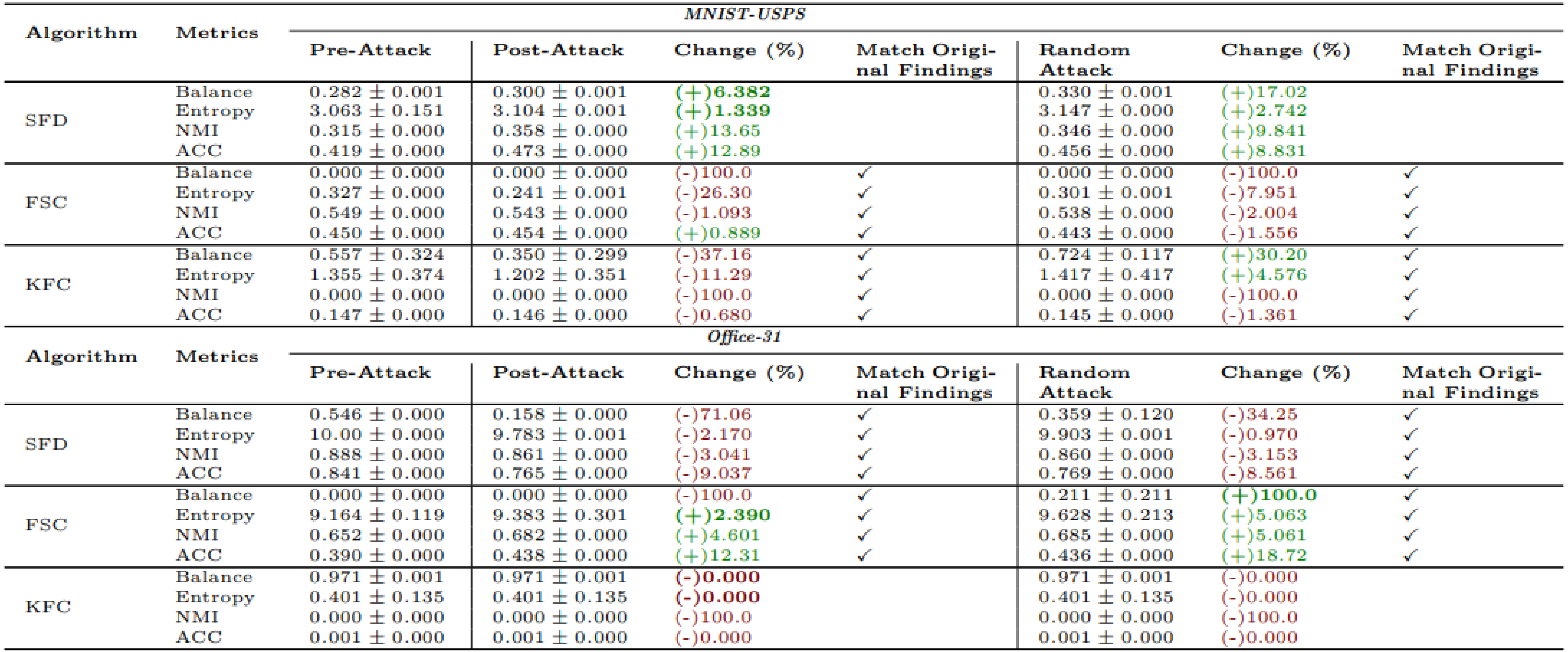
Claim 2: Fair K-Center, Fair Spectral Clustering, and Scalable Fairlet Decomposition, demonstrate a lack of robustness to adversarial influence, exhibiting significant volatility in terms of fairness utility metrics such as Balance and Entropy.

→ REPRODUCED

Claim 3: *Consensus Fair Clustering* exhibits high resilience against the proposed fairness attack, offering a robust solution for achieving fair clustering.

→ REPRODUCED

Claim 1: Partially Reproduced



MNIST-USPS

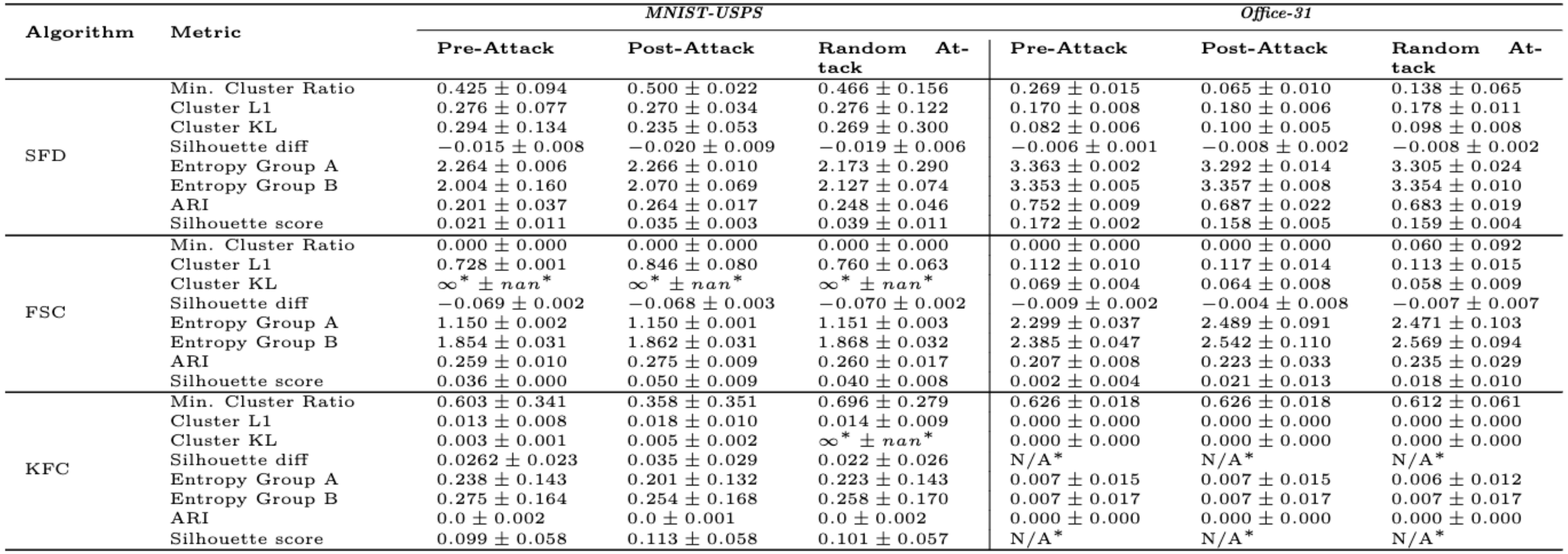
Algorithm	Metrics	Pre-Attack	Post-Attack	Change (%)	Match Original Findings	Random Attack	Change (%)	Match Original Findings
SFD	Balance	0.282 ± 0.001	0.300 ± 0.001	**(+)6.382**		0.330 ± 0.001	(+)17.02	
	Entropy	3.063 ± 0.151	3.104 ± 0.001	**(+)1.339**		3.147 ± 0.000	(+)2.742	
	NMI	0.315 ± 0.000	0.358 ± 0.000	(+)13.65		0.346 ± 0.000	(+)9.841	
	ACC	0.419 ± 0.000	0.473 ± 0.000	(+)12.89		0.456 ± 0.000	(+)8.831	
FSC	Balance	0.000 ± 0.000	0.000 ± 0.000	(-)100.0	✓	0.000 ± 0.000	(-)100.0	✓
	Entropy	0.327 ± 0.000	0.241 ± 0.001	(-)26.30	✓	0.301 ± 0.001	(-)7.951	✓
	NMI	0.549 ± 0.000	0.543 ± 0.000	(-)1.093	✓	0.538 ± 0.000	(-)2.004	✓
	ACC	0.450 ± 0.000	0.454 ± 0.000	(+)0.889	✓	0.443 ± 0.000	(-)1.556	✓
KFC	Balance	0.557 ± 0.324	0.350 ± 0.299	(-)37.16	✓	0.724 ± 0.117	(+)30.20	✓
	Entropy	1.355 ± 0.374	1.202 ± 0.351	(-)11.29	✓	1.417 ± 0.417	(+)4.576	✓
	NMI	0.000 ± 0.000	0.000 ± 0.000	(-)100.0	✓	0.000 ± 0.000	(-)100.0	✓
	ACC	0.147 ± 0.000	0.146 ± 0.000	(-)0.680	✓	0.145 ± 0.000	(-)1.361	✓

Office-31

Algorithm	Metrics	Pre-Attack	Post-Attack	Change (%)	Match Original Findings	Random Attack	Change (%)	Match Original Findings
SFD	Balance	0.546 ± 0.000	0.158 ± 0.000	(-)71.06	✓	0.359 ± 0.120	(-)34.25	✓
	Entropy	10.00 ± 0.000	9.783 ± 0.001	(-)2.170	✓	9.903 ± 0.001	(-)0.970	✓
	NMI	0.888 ± 0.000	0.861 ± 0.000	(-)3.041	✓	0.860 ± 0.000	(-)3.153	✓
	ACC	0.841 ± 0.000	0.765 ± 0.000	(-)9.037	✓	0.769 ± 0.000	(-)8.561	✓
FSC	Balance	0.000 ± 0.000	0.000 ± 0.000	(-)100.0	✓	0.211 ± 0.211	**(+)100.0**	✓
	Entropy	9.164 ± 0.119	9.383 ± 0.301	**(+)2.390**	✓	9.628 ± 0.213	(+)5.063	✓
	NMI	0.652 ± 0.000	0.682 ± 0.000	(+)4.601	✓	0.685 ± 0.000	(+)5.061	✓
	ACC	0.390 ± 0.000	0.438 ± 0.000	(+)12.31	✓	0.436 ± 0.000	(+)18.72	✓
KFC	Balance	0.971 ± 0.001	0.971 ± 0.001	**(-)0.000**		0.971 ± 0.001	(-)0.000	
	Entropy	0.401 ± 0.135	0.401 ± 0.135	**(-)0.000**		0.401 ± 0.135	(-)0.000	
	NMI	0.000 ± 0.000	0.000 ± 0.000	(-)100.0		0.000 ± 0.000	(-)100.0	
	ACC	0.001 ± 0.000	0.001 ± 0.000	(-)0.000		0.001 ± 0.000	(-)0.000	

Results beyond original paper

Additional Metrics

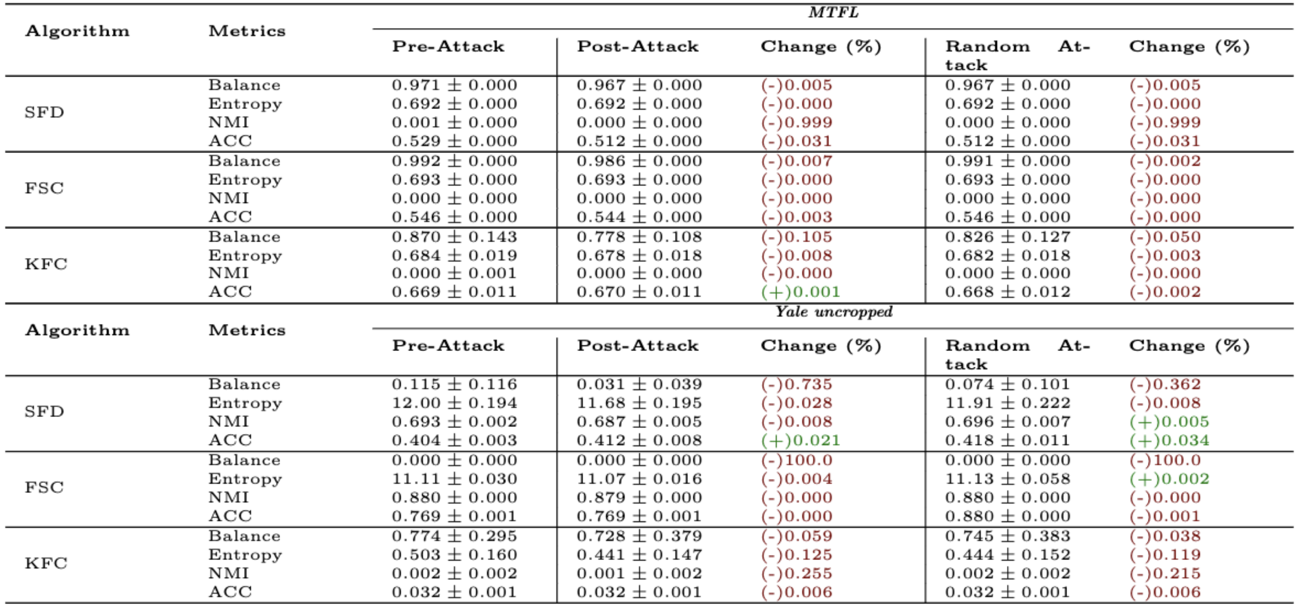


Algorithm	Metric	*MNIST-USPS*			*Office-31*		
		Pre-Attack	**Post-Attack**	**Random Attack**	**Pre-Attack**	**Post-Attack**	**Random Attack**
SFD	Min. Cluster Ratio	0.425 ± 0.094	0.500 ± 0.022	0.466 ± 0.156	0.269 ± 0.015	0.065 ± 0.010	0.138 ± 0.065
	Cluster L1	0.276 ± 0.077	0.270 ± 0.034	0.276 ± 0.122	0.170 ± 0.008	0.180 ± 0.006	0.178 ± 0.011
	Cluster KL	0.294 ± 0.134	0.235 ± 0.053	0.269 ± 0.300	0.082 ± 0.006	0.100 ± 0.005	0.098 ± 0.008
	Silhouette diff	−0.015 ± 0.008	−0.020 ± 0.009	−0.019 ± 0.006	−0.006 ± 0.001	−0.008 ± 0.002	−0.008 ± 0.002
	Entropy Group A	2.264 ± 0.006	2.266 ± 0.010	2.173 ± 0.290	3.363 ± 0.002	3.292 ± 0.014	3.305 ± 0.024
	Entropy Group B	2.004 ± 0.160	2.070 ± 0.069	2.127 ± 0.074	3.353 ± 0.005	3.357 ± 0.008	3.354 ± 0.010
	ARI	0.201 ± 0.037	0.264 ± 0.017	0.248 ± 0.046	0.752 ± 0.009	0.687 ± 0.022	0.683 ± 0.019
	Silhouette score	0.021 ± 0.011	0.035 ± 0.003	0.039 ± 0.011	0.172 ± 0.002	0.158 ± 0.005	0.159 ± 0.004
FSC	Min. Cluster Ratio	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.060 ± 0.092
	Cluster L1	0.728 ± 0.001	0.846 ± 0.080	0.760 ± 0.063	0.112 ± 0.010	0.117 ± 0.014	0.113 ± 0.015
	Cluster KL	$\infty^* \pm nan^*$	$\infty^* \pm nan^*$	$\infty^* \pm nan^*$	0.069 ± 0.004	0.064 ± 0.008	0.058 ± 0.009
	Silhouette diff	−0.069 ± 0.002	−0.068 ± 0.003	−0.070 ± 0.002	−0.009 ± 0.002	−0.004 ± 0.008	−0.007 ± 0.007
	Entropy Group A	1.150 ± 0.002	1.150 ± 0.001	1.151 ± 0.003	2.299 ± 0.037	2.489 ± 0.091	2.471 ± 0.103
	Entropy Group B	1.854 ± 0.031	1.862 ± 0.031	1.868 ± 0.032	2.385 ± 0.047	2.542 ± 0.110	2.569 ± 0.094
	ARI	0.259 ± 0.010	0.275 ± 0.009	0.260 ± 0.017	0.207 ± 0.008	0.223 ± 0.033	0.235 ± 0.029
	Silhouette score	0.036 ± 0.000	0.050 ± 0.009	0.040 ± 0.008	0.002 ± 0.004	0.021 ± 0.013	0.018 ± 0.010
KFC	Min. Cluster Ratio	0.603 ± 0.341	0.358 ± 0.351	0.696 ± 0.279	0.626 ± 0.018	0.626 ± 0.018	0.612 ± 0.061
	Cluster L1	0.013 ± 0.008	0.018 ± 0.010	0.014 ± 0.009	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Cluster KL	0.003 ± 0.001	0.005 ± 0.002	$\infty^* \pm nan^*$	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Silhouette diff	0.0262 ± 0.023	0.035 ± 0.029	0.022 ± 0.026	N/A*	N/A*	N/A*
	Entropy Group A	0.238 ± 0.143	0.201 ± 0.132	0.223 ± 0.143	0.007 ± 0.015	0.007 ± 0.015	0.006 ± 0.012
	Entropy Group B	0.275 ± 0.164	0.254 ± 0.168	0.258 ± 0.170	0.007 ± 0.017	0.007 ± 0.017	0.007 ± 0.017
	ARI	0.0 ± 0.002	0.0 ± 0.001	0.0 ± 0.002	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
	Silhouette score	0.099 ± 0.058	0.113 ± 0.058	0.101 ± 0.057	N/A*	N/A*	N/A*

Table 4: Results for pre-attack, post-attack (*black-box*), and random attack, when 15% group membership labels are switched for fair clustering algorithms SFD, FSC, and KFC and datasets *MNIST-USPS* and *Office-31*. Results show the impact on additional metrics, where N/A corresponds to uniform clustering, ∞ to infinite values, and *nan* to undefined values.

- Form a better picture of the performance of the model
- Provide insights to design new attacks

Additional Datasets



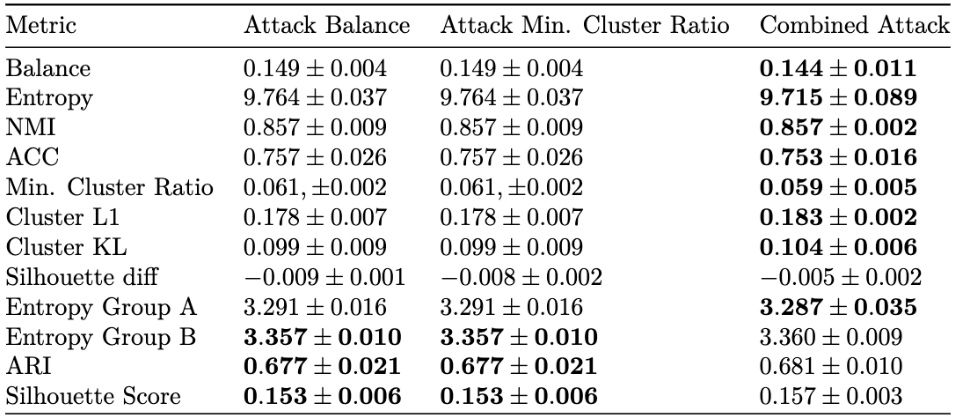
Algorithm	Metrics	*MTFL*				
		Pre-Attack	Post-Attack	Change (%)	Random Attack	Change (%)
SFD	Balance	0.971 ± 0.000	0.967 ± 0.000	(-)0.005	0.967 ± 0.000	(-)0.005
	Entropy	0.692 ± 0.000	0.692 ± 0.000	(-)0.000	0.692 ± 0.000	(-)0.000
	NMI	0.001 ± 0.000	0.000 ± 0.000	(-)0.999	0.000 ± 0.000	(-)0.999
	ACC	0.529 ± 0.000	0.512 ± 0.000	(-)0.031	0.512 ± 0.000	(-)0.031
FSC	Balance	0.992 ± 0.000	0.986 ± 0.000	(-)0.007	0.991 ± 0.000	(-)0.002
	Entropy	0.693 ± 0.000	0.693 ± 0.000	(-)0.000	0.693 ± 0.000	(-)0.000
	NMI	0.000 ± 0.000	0.000 ± 0.000	(-)0.000	0.000 ± 0.000	(-)0.000
	ACC	0.546 ± 0.000	0.544 ± 0.000	(-)0.003	0.546 ± 0.000	(-)0.000
KFC	Balance	0.870 ± 0.143	0.778 ± 0.108	(-)0.105	0.826 ± 0.127	(-)0.050
	Entropy	0.684 ± 0.019	0.678 ± 0.018	(-)0.008	0.682 ± 0.018	(-)0.003
	NMI	0.000 ± 0.001	0.000 ± 0.000	(-)0.000	0.000 ± 0.000	(-)0.000
	ACC	0.669 ± 0.011	0.670 ± 0.011	(+)0.001	0.668 ± 0.012	(-)0.002

Algorithm	Metrics	*Yale uncropped*				
		Pre-Attack	Post-Attack	Change (%)	Random Attack	Change (%)
SFD	Balance	0.115 ± 0.116	0.031 ± 0.039	(-)0.735	0.074 ± 0.101	(-)0.362
	Entropy	12.00 ± 0.194	11.68 ± 0.195	(-)0.028	11.91 ± 0.222	(-)0.008
	NMI	0.693 ± 0.002	0.687 ± 0.005	(-)0.008	0.696 ± 0.007	(+)0.005
	ACC	0.404 ± 0.003	0.412 ± 0.008	(+)0.021	0.418 ± 0.011	(+)0.034
FSC	Balance	0.000 ± 0.000	0.000 ± 0.000	(-)100.0	0.000 ± 0.000	(-)100.0
	Entropy	11.11 ± 0.030	11.07 ± 0.016	(-)0.004	11.13 ± 0.058	(+)0.002
	NMI	0.880 ± 0.000	0.879 ± 0.000	(-)0.000	0.880 ± 0.000	(-)0.000
	ACC	0.769 ± 0.001	0.769 ± 0.001	(-)0.000	0.880 ± 0.000	(-)0.001
KFC	Balance	0.774 ± 0.295	0.728 ± 0.379	(-)0.059	0.745 ± 0.383	(-)0.038
	Entropy	0.503 ± 0.160	0.441 ± 0.147	(-)0.125	0.444 ± 0.152	(-)0.119
	NMI	0.002 ± 0.002	0.001 ± 0.002	(-)0.255	0.002 ± 0.002	(-)0.215
	ACC	0.032 ± 0.001	0.032 ± 0.001	(-)0.006	0.032 ± 0.001	(-)0.006

Table 5: Results for pre-attack, post-attack (*black-box*), and random attack, when 15% group membership labels are switched for fair clustering algorithms SFD, FSC, and KFC and datasets *MTFL* and *Yale uncropped*. Results show the impact on fairness utility (Balance and Entropy) and clustering utility (NMI and ACC). Relative changes provide insights into how our changes between pre-attack and post-attack / random attack differ from those of the paper.

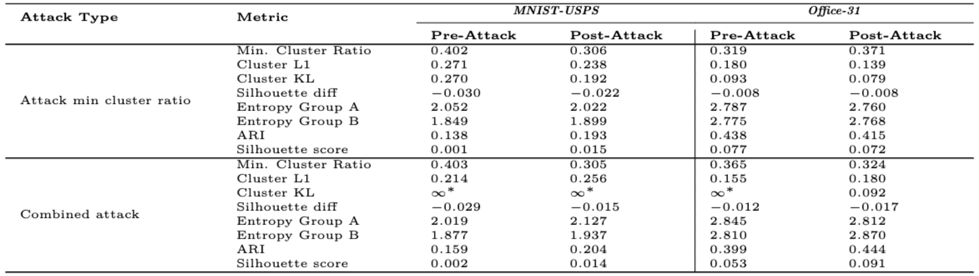
- Contribute to evaluating the performance of the models in other data distributions
- Still indicate that CFC performs better than other fairness algorithms

Additional Attack Methods



Metric	Attack Balance	Attack Min. Cluster Ratio	Combined Attack
Balance	0.149 ± 0.004	0.149 ± 0.004	$\mathbf{0.144 \pm 0.011}$
Entropy	9.764 ± 0.037	9.764 ± 0.037	$\mathbf{9.715 \pm 0.089}$
NMI	0.857 ± 0.009	0.857 ± 0.009	$\mathbf{0.857 \pm 0.002}$
ACC	0.757 ± 0.026	0.757 ± 0.026	$\mathbf{0.753 \pm 0.016}$
Min. Cluster Ratio	$0.061, \pm 0.002$	$0.061, \pm 0.002$	$\mathbf{0.059 \pm 0.005}$
Cluster L1	0.178 ± 0.007	0.178 ± 0.007	$\mathbf{0.183 \pm 0.002}$
Cluster KL	0.099 ± 0.009	0.099 ± 0.009	$\mathbf{0.104 \pm 0.006}$
Silhouette diff	-0.009 ± 0.001	-0.008 ± 0.002	-0.005 ± 0.002
Entropy Group A	3.291 ± 0.016	3.291 ± 0.016	$\mathbf{3.287 \pm 0.035}$
Entropy Group B	$\mathbf{3.357 \pm 0.010}$	$\mathbf{3.357 \pm 0.010}$	3.360 ± 0.009
ARI	$\mathbf{0.677 \pm 0.021}$	$\mathbf{0.677 \pm 0.021}$	0.681 ± 0.010
Silhouette Score	$\mathbf{0.153 \pm 0.006}$	$\mathbf{0.153 \pm 0.006}$	0.157 ± 0.003

Table 10: Results of Additional Attack Methods: This table compares the performance of the original balance attack against the newly introduced Minimum Cluster Ratio attack and the Combined (Balance & Entropy) attack. For a consistent comparison, the results presented are based on the same three seeds that were utilized during the grid search. The most effective attack strategy for each metric, as indicated by the lowest value, is emphasized in bold.



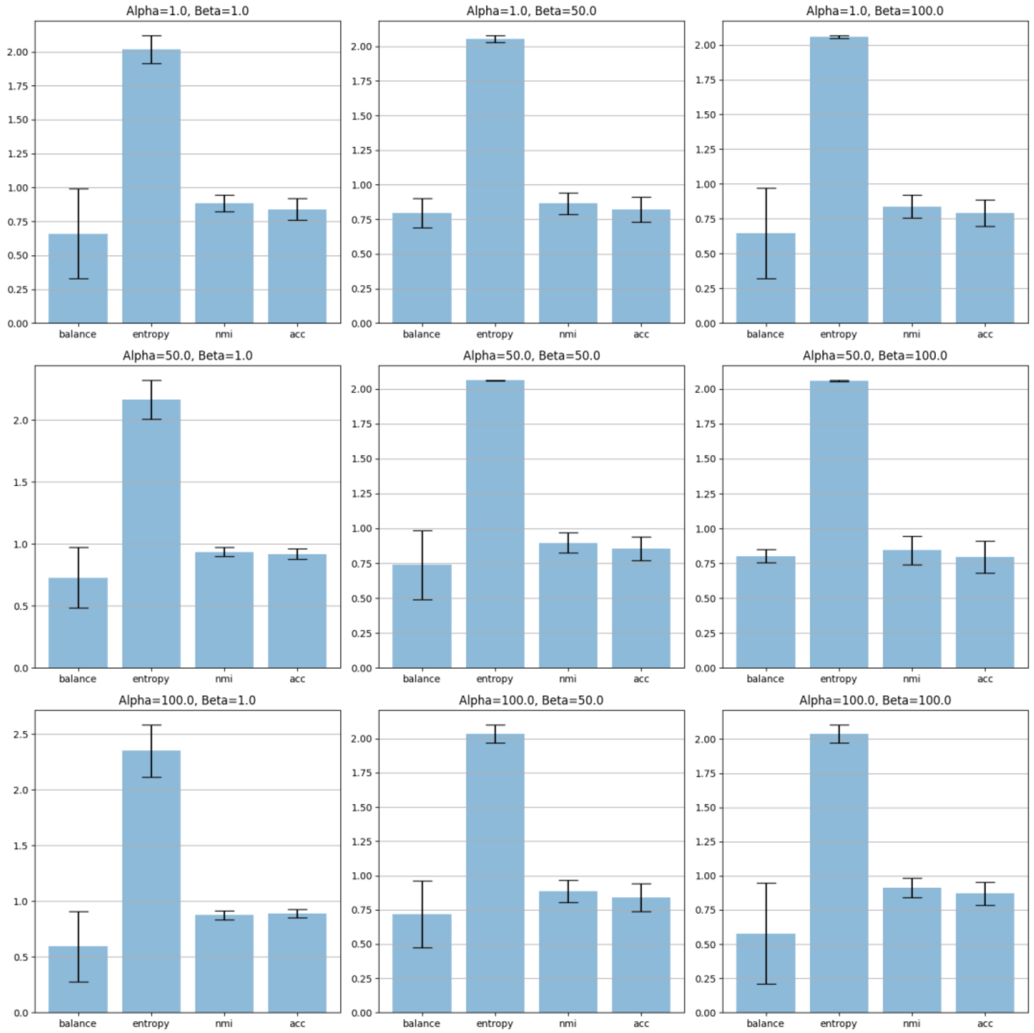
Attack Type	**Metric**	*MNIST-USPS*		*Office-31*	
		Pre-Attack	**Post-Attack**	**Pre-Attack**	**Post-Attack**
Attack min cluster ratio	Min. Cluster Ratio	0.402	0.306	0.319	0.371
	Cluster L1	0.271	0.238	0.180	0.139
	Cluster KL	0.270	0.192	0.093	0.079
	Silhouette diff	−0.030	−0.022	−0.008	−0.008
	Entropy Group A	2.052	2.022	2.787	2.760
	Entropy Group B	1.849	1.899	2.775	2.768
	ARI	0.138	0.193	0.438	0.415
	Silhouette score	0.001	0.015	0.077	0.072
Combined attack	Min. Cluster Ratio	0.403	0.305	0.365	0.324
	Cluster L1	0.214	0.256	0.155	0.180
	Cluster KL	∞^*	∞^*	∞^*	0.092
	Silhouette diff	−0.029	−0.015	−0.012	−0.017
	Entropy Group A	2.019	2.127	2.845	2.812
	Entropy Group B	1.877	1.937	2.810	2.870
	ARI	0.159	0.204	0.399	0.444
	Silhouette score	0.002	0.014	0.053	0.091

Table 11: Results for pre-attack, post-attack, when 15% group membership labels are switched for defense algorithm CFC and datasets *MNIST-USPS* and *Office-31*. Attack types are attack on Minimum Cluster Ratio and Combined Attack. Experiments are run once because of the consumed GPU hours and the small standard deviations in other related experiments.

- Improvements observed in attack performance with this configuration, as compared to the originally proposed attack, were marginal

- CFC algorithm demonstrates robustness in countering these attacks

Ablation Study



- Focused on two key hyperparameters alpha (α) and beta (β)
- Minimal influence of the α and β hyperparameters on the CFC model

Discussion

- **Most** results that were provided in the original Robust Fair Clustering paper were **reproduced**.
- Overall CFC had **superior** performance and was more **robust**.

Limitations and Future Work

- Limited computational resources and time
- Expand the grid search
- More effective attack strategy

Thank you for your attention!