Reproducing “Robust Fair Clustering: A Novel Fairness Attack and Defense Framework"

Lucas Ponticelli, Vincent Loos, Eren Kocadag, Kacper Bartosik
University of Amsterdam

Overview

The objective of our paper was to verify the **results** of **Chhabra et al.’s** work on **fairness** in clustering under **adversarial attacks**. [1] This is of importance, as **clustering** algorithms may impact sensitive domains like finance or criminal justice. [2]

The original authors made three main claims in their paper:

- **State-of-the-art fair clustering models** are highly susceptible to **adversarial attacks**, which can significantly diminish their **fairness performance**.
- The **novel black-box attack** effectively degrades the **fairness performance** by altering a small portion of protected group memberships.
- The proposed **Consensus Fair Clustering (CFC)** defense mechanism not only resists **adversarial attacks** but can also maintain or even improve **clustering performance** post-attack.

Scope of Reproducibility

Our aim was the validate the claims made by **Chhabra et al.** and to test them on **new datasets**. The original paper is also **extended** by attacking both **fairness metrics** evaluated in our paper (**Balance** and **Entropy**), and measuring the impact of each attack on the other metric.

Datasets

The original study used the **MNIST-USPS**, **Office-31**, **Digits**, and **Yale** datasets. We extended this by introducing the **FairFace**, **OULAD**, and **Dutch Census** datasets. This introduces real-world sensitive features and tabular data.

The Attack

Three fair clustering approaches were considered:

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

The attack can be considered a **minimisation problem**:

$$min_{G_A} \phi(\theta(O, G_D), G_D) \text{ s.t. } O = F(X, K, (G_A, G_D))$$

Where a subgroup of sensitive data, denoted as $\mathbf{G_A}$, is perturbed such that a **fairness metric** (**Entropy** or **Balance**), represented by $\boldsymbol{\phi}$, is minimized. For clustering performance, an **unsupervised equivalent of accuracy** and the **Normalized Mutual Information (NMI)** metrics are considered.

The Defense

The **CFC mechanism** uses consensus clustering with **fairness constraints**, transforming the task into **graph partitioning**. It leverages a **novel graph-based neural network architecture** for learning representations tailored to **fair clustering**.

- In **stage one,** a **co-association matrix** is learned.
- In **stage two, graph embeddings** are created for **fair clustering**.

This approach ensures **robustness** against **adversarial attacks**. A complete overview of the mechanism is showcased in **Figure 1**.

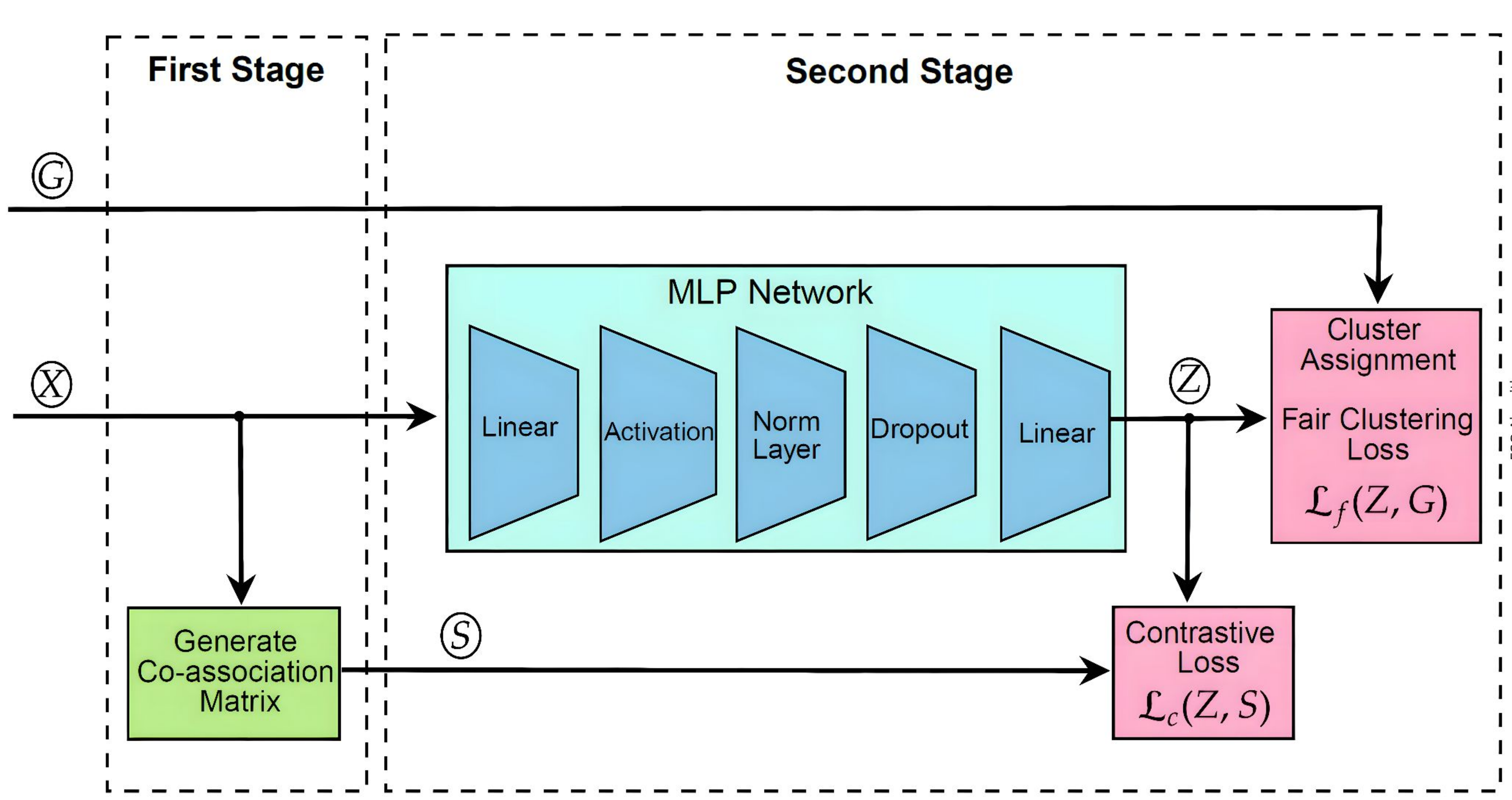

Figure 1: Proposed CFC framework. Adopted from Chhabra et al. (2023). [1]

Their attack

Our attack

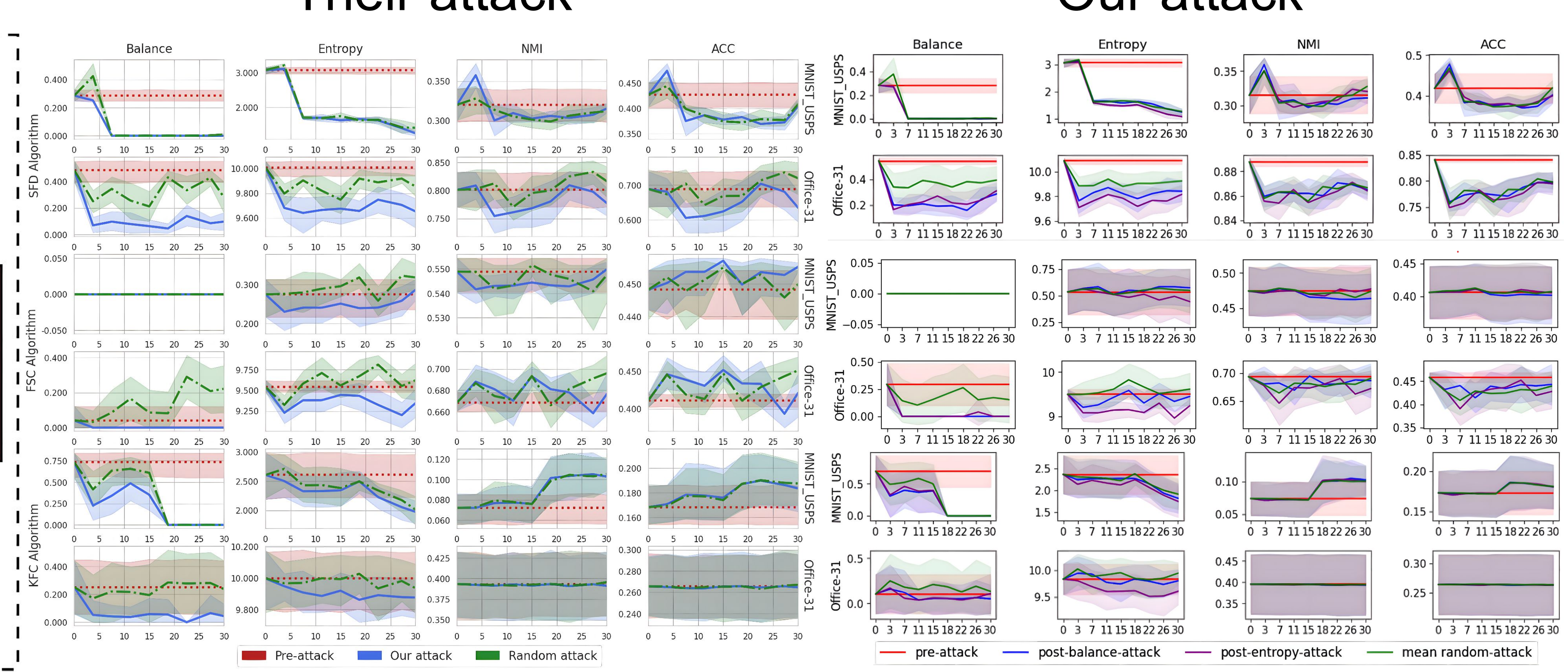

Figure 2: Comparison of original and reproduced attack results, showing alignment and the effectiveness of targeted attacks in degrading fairness.

Their defense

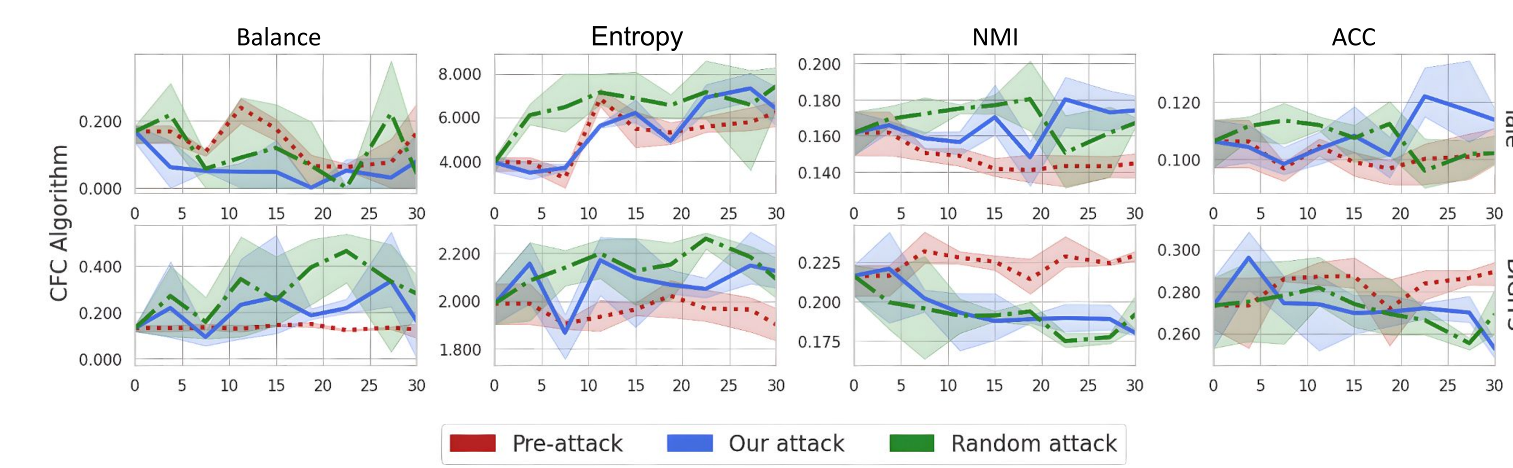

Figure 3: Original CFC defense results demonstrating robustness against adversarial attacks.

Our defense

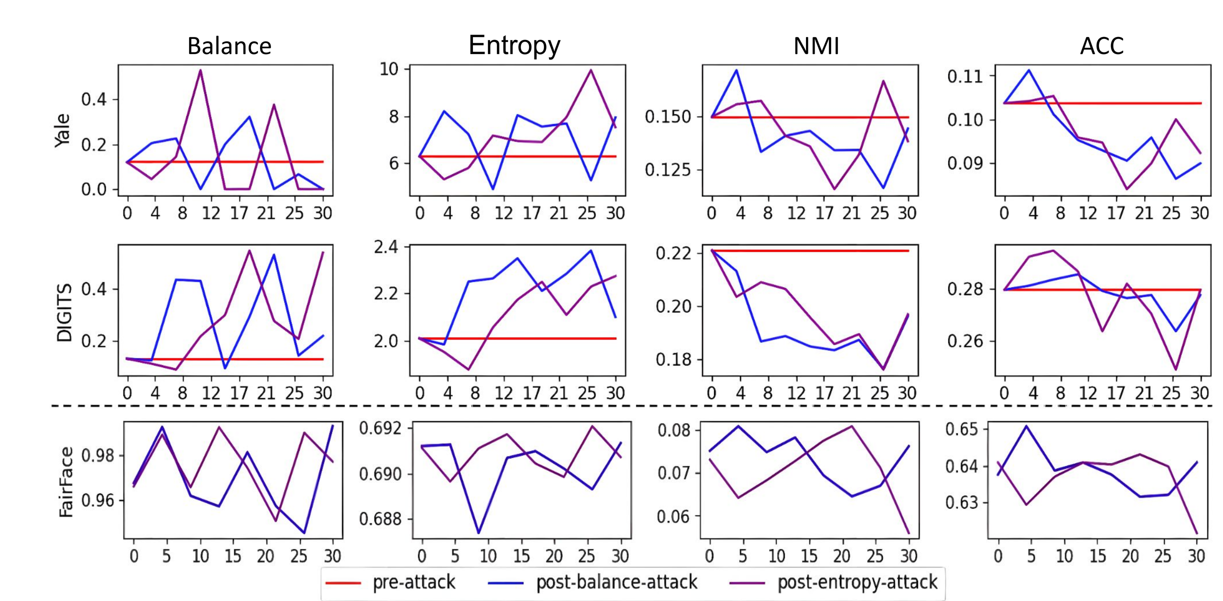

Figure 4: Reproduced CFC defense results confirming robustness against adversarial attacks across datasets.

Extensions

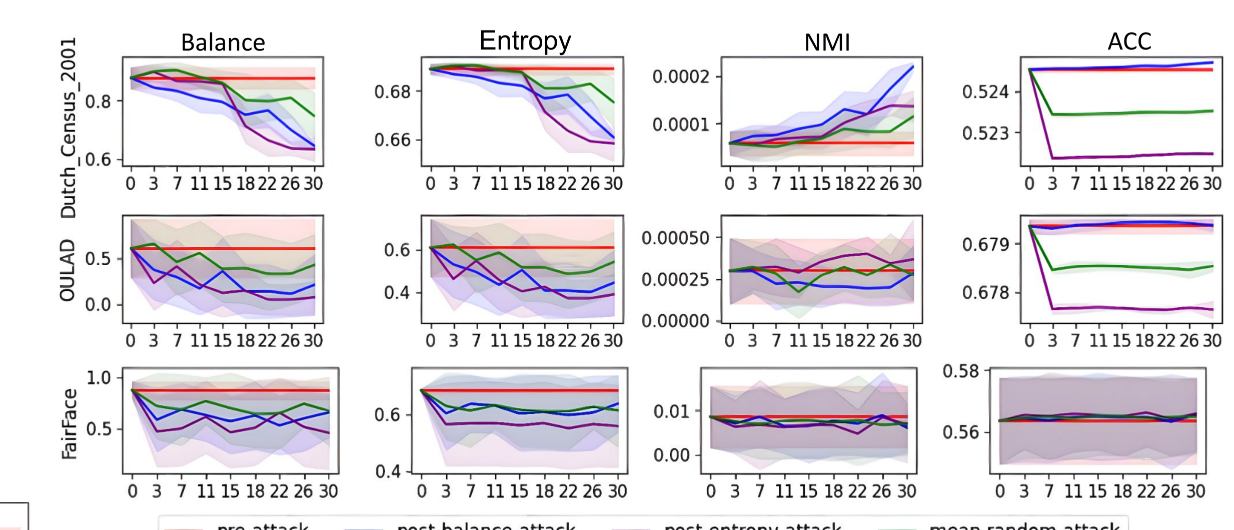

Figure 5: Extension of KFC algorithm experiments to new datasets (FairFace, OULAD, Dutch Census).

Conclusion

We reproduced and extended the original experiments, confirming the **vulnerability** of fair clustering models to **adversarial attacks** and the **robustness** of the CFC defense, as outlined in the overview.

References

[1] Chhabra A. et al. (2023). “Robust Fair Clustering: A Novel Fairness Attack and Defense Framework.” In: The Eleventh International Conference on Learning Representations

[2] Ghosal A. et al. (2020). “A short review on different clustering techniques and their applications.” In: Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, pp. 69–83.