Taxonomy of Fairness Concepts in Clustering

Fair clustering has been a very active and prolific area of research in the last few years, with a plethora of fairness notions studied. A crucial part of this tutorial focuses on creating a rigorous taxonomy that provides a thorough categorization of all the different concepts of fairness in clustering. In what follows, we give a brief description of this taxonomy.

Notions of Fairness in Clustering:

1. **Demographic Fairness:** The high-level goal of such notions is to treat each group of points fairly, with respect to either how other groups are being treated or with respect to the specific needs of the group at hand. Usually, the groups of points that we are interested in are given as part of the input, and they represent points sharing some important defining attributes. Hence, each group can be perceived as an underlying demographic of the population. For example, we might be provided with groups representing age, e.g., "children", "teenagers", "adults", "senior citizens". We now present all different concepts of demographic fairness that have been studied in the context of clustering.

   (a) **Balanced Representation:** This is the first fair clustering problem ever studied, and it was introduced in the seminal work of [13] for the setting of metric k-clustering. It asks for clusters in which for every demographic, the ratio of its points in every cluster is the same as the ratio of its points over the whole dataset. This notion has been further studied in the k-clustering setting [8, 30], as well as in other clustering paradigms such as spectral clustering [25] and correlation clustering [3].

   (b) **Relaxed Notion of Balance:** Here we do not require the ratio of the points of each demographic in the cluster to be exactly the same as the ratio of its points over the whole dataset. We rather relax this condition, and ask for a ratio that is within a certain input given range. This concept was simultaneously introduced by [8, 7] in the context of k-clustering. It was later studied for hierarchical clustering by [2], in a stochastic setting by [16], and with a fairness maximization objective by [15].

   (c) **Fairness in Center Selection**: In the context of k-clustering where k cluster centers need to be selected, it is reasonable to impose fairness constraints on the set of selected centers. The work of [24, 21] focuses on avoiding over-representation of any demographic group in the set of selected centers, while [31] focuses on avoiding under-representation of any group in the chosen centers.

   (d) **Proportionally Fair Clustering:** In this model, we assume that every large enough group of points is entitled to their own cluster center. Therefore, we should make sure that for every such group of points, their assigned centers are close enough to all of them, in the sense that there exists no other point that is closer to every point in the coalition, thus giving them an incentive to deviate from the current assignment. An interesting aspect of this model is that there are no a priori given sets of points, and the fairness guarantee should simultaneously hold for every possible subset of the dataset. This setting was introduced by [12], with further refined results given by [28].

(e) **Demographic Fairness in Clustering with Outliers:** When clustering with outliers, we are allowed to exclude a certain number of points from our computations, thus leaving them "unclustered". Being chosen as an outlier is an inherently disadvantageous event for a point. Hence, when the dataset involves points coming from different demographic groups, we should make sure that a fair amount of points from each such group is chosen as outliers. This model has been studied only under the k-center objective, with [6] being the paper that introduced it. Later on, [5, 20] gave improved results for it.

(f) **Socially Fair k-Clustering:** In many clustering applications, the quantity that really matters to each point is the distance to its assigned center. Hence, in the presence of multiple demographic groups it makes sense to consider a fairness metric that looks at the average assignment distance of each demographic. [1, 17] independently introduced a model where the objective is to minimize the maximum average assignment distance over all demographic groups. Later on, [27, 18] gave improved results for the problem.

2. **Individual Fairness:** The high-level goal of these notions is to treat each individual point fairly, with respect to either how other individual points are being treated or with respect to the specific needs of the point at hand. Therefore, our attention here shifts from groups to individuals. As we will argue during the tutorial, the demographic concepts of fairness mentioned earlier fail to capture the individualistic needs of points, thus making the need for studying individual fairness more vital.

   The seminal work of [14] will help us create a taxonomy for concepts of individual fairness. Specifically, [14] introduced a notion of individual fairness for classification, where the aim was to **treat similar individuals similarly**. Our categorization makes a distinction between notions of fairness for clustering that follow the paradigm of [14], and notions that do not.

   (a) **Notions of Individual Fairness that follow the paradigm of [14]:** In order to translate the paradigm of [14] to a well-defined clustering problem, one needs to answer two questions. The first is how can we define similarity between points in the context of clustering? The second is what does similar treatment mean in a clustering problem? The first question is a modeling issue and hence not that important. The second one is far more crucial, since the answer to it defines what we perceive as fair treatment. Therefore, we proceed with the following categorization.

      i. **Similar treatment in the sense of receiving the same cluster placement:** Here, points that are deemed similar should eventually be placed in the same cluster. The papers of [9, 10, 4] follow this concept, by guaranteeing the same cluster placement in some stochastic way.

      ii. **Similar treatment in the sense of receiving similar assignment distances:** As we mentioned earlier, in many clustering applications, the quantity that really matters to each point is the distance to its assigned center. The work of [11] tries to provide comparable assignment distances to similar points.

   (b) **Notions of Individual Fairness that diverge from the paradigm of [14]:**

      i. **A Center in My Neighborhood:** In certain applications, minimizing a global objective on the assignment distances, e.g., the k-means objective, does not suffice in order to capture the special needs of individual points. For instance, some points may require an assignment distance much smaller than what other points can tolerate. Hence, it is reasonable to seek solutions where we do not only minimize the global objective, but also try to satisfy the individual distance needs of every point. This model was introduced in [22], and increasingly better results were given in a subsequent series of papers [26, 29, 32].

ii. **Individual Fairness when clustering with outliers:** As discussed earlier, being chosen as an outlier is an inherently disadvantageous event for a point. Therefore, a solution that consistently picks certain points as outliers can be arguably considered as biased against them. [19] introduced a randomized model, where each point is not picked as an outlier with probability at least a certain value. Hence, in a stochastic sense this type of solution guarantees that every individual point has a decent chance of getting clustered. Further, [5] gave better results for this problem.

iii. **Proximity to the points of your cluster:** Finally, [23] introduce an intriguing concept, that views individual fairness as ensuring that each point is on average closer to the points in its own cluster than to the points in any other cluster.

References

[1] Mohsen Abbasi, Aditya Bhaskara, and Suresh Venkatasubramanian. "Fair Clustering via Equitable Group Representations". In: FAccT '21. 2021.

[2] Sara Ahmadian, Alessandro Epasto, Marina Knittel, Ravi Kumar, Mohammad Mahdian, Benjamin Moseley, Philip Pham, Sergei Vassilvitskii, and Yuyan Wang. "Fair hierarchical clustering". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 21050–21060.

[3] Sara Ahmadian, Alessandro Epasto, Ravi Kumar, and Mohammad Mahdian. "Fair correlation clustering". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 4195–4205.

[4] Nihesh Anderson, Suman K. Bera, Syamantak Das, and Yang Liu. "Distributional Individual Fairness in Clustering". In: *CoRR* abs/2006.12589 (2020). URL: https://arxiv.org/abs/2006.12589.

[5] Georg Anegg, Haris Angelidakis, Adam Kurpisz, and Rico Zenklusen. "A Technique for Obtaining True Approximations for k-Center with Covering Constraints". In: *Integer Programming and Combinatorial Optimization*. Ed. by Daniel Bienstock and Giacomo Zambelli. Cham: Springer International Publishing, 2020, pp. 52–65.

[6] Sayan Bandyapadhyay, Tanmay Inamdar, Shreyas Pai, and Kasturi Varadarajan. "A Constant Approximation for Colorful k-Center". In: *27th Annual European Symposium on Algorithms (ESA 2019)*. Vol. 144. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2019, 12:1–12:14.

[7] Suman Bera, Deeparnab Chakrabarty, Nicolas Flores, and Maryam Negahbani. "Fair Algorithms for Clustering". In: *Advances in Neural Information Processing Systems 32*. 2019, pp. 4954–4965.

[8] Ioana O. Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R. Schmidt, and Melanie Schmidt. "On the Cost of Essentially Fair Clusterings". In: *APPROX/RANDOM 2019*. Vol. 145. 2019, 18:1–18:22.

[9] Brian Brubach, Darshan Chakrabarti, John P. Dickerson, Samir Khuller, Aravind Srinivasan, and Leonidas Tsepenekas. "A Pairwise Fair and Community-preserving Approach to k-Center Clustering". In: *ICML*. 2020.

[10] Brian Brubach, Darshan Chakrabarti, John P. Dickerson, Aravind Srinivasan, and Leonidas Tsepenekas. "Fairness, Semi-Supervised Learning, and More: A General Framework for Clustering with Stochastic Pairwise Constraints". In: *AAAI*. 2021.

[11] Darshan Chakrabarti, Seyed A. Esmaeili, John P. Dickerson, Aravind Srinivasan, and Leonidas Tsepenekas. "A New Notion of Individually Fair Clustering: α-Equitable k-Center". In: *AISTATS*. 2022.

[12] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. "Proportionally Fair Clustering". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 1032–1041.

[13] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. "Fair Clustering Through Fairlets". In: *Advances in Neural Information Processing Systems 30*. 2017.

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness Through Awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. 2012.

[15] Seyed A. Esmaeili, Brian Brubach, Aravind Srinivasan, and John P. Dickerson. "Fair Clustering Under a Bounded Cost". In: *NeurIPS*. 2021.

[16] Seyed A. Esmaeili, Brian Brubach, Leonidas Tsepenekas, and John P. Dickerson. "Probabilistic Fair Clustering". In: *NeurIPS*. 2020.

[17] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. "Socially Fair K-Means Clustering". In: FAccT '21. 2021.

[18] Dishant Goyal and Ragesh Jaiswal. *Tight FPT Approximation for Socially Fair Clustering*. 2021. arXiv: 2106.06755 [cs.DS].

[19] David G. Harris, Thomas Pensyl, Aravind Srinivasan, and Khoa Trinh. "A Lottery Model for Center-Type Problems With Outliers". In: *ACM Trans. Algorithms* 15.3 (June 2019). ISSN: 1549-6325. DOI: 10.1145/3311953. URL: https://doi.org/10.1145/3311953.

[20] Xinrui Jia, Ksheteej Sheth, and Ola Svensson. "Fair colorful k-center clustering". In: *Mathematical Programming* (2021), pp. 1–22.

[21] Matthew Jones, Huy Nguyen, and Thy Nguyen. "Fair k-centers via maximum matching". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 4940–4949.

[22] Christopher Jung, Sampath Kannan, and Neil Lutz. *A Center in Your Neighborhood: Fairness in Facility Location*. 2019. arXiv: 1908.09041 [cs.DS].

[23] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. *A Notion of Individual Fairness for Clustering*. 2020. arXiv: 2006.04960 [stat.ML].

[24] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. "Fair k-Center Clustering for Data Summarization". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 3448–3457.

[25] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. "Guarantees for spectral clustering with fairness constraints". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 3458–3467.

[26] Sepideh Mahabadi and Ali Vakilian. "Individual Fairness for k-Clustering". In: *Proceedings of the 37th International Conference on Machine Learning*. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 6586–6596.

[27] Yury Makarychev and Ali Vakilian. "Approximation Algorithms for Socially Fair Clustering". In: *Proceedings of Thirty Fourth Conference on Learning Theory*. Vol. 134. Proceedings of Machine Learning Research. PMLR, 15–19 Aug 2021, pp. 3246–3264.

[28] Evi Micha and Nisarg Shah. "Proportionally Fair Clustering Revisited". In: *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*. Vol. 168. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020, 85:1–85:16. ISBN: 978-3-95977-138-2.

[29] Maryam Negahbani and Deeparnab Chakrabarty. "Better Algorithms for Individually Fair k-Clustering". In: *Advances in Neural Information Processing Systems* 34 (2021).

[30] Clemens Rösner and Melanie Schmidt. "Privacy Preserving Clustering with Constraints". In: *45th International Colloquium on Automata, Languages, and Programming (ICALP 2018)*. Vol. 107. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, 96:1–96:14. ISBN: 978-3-95977-076-7.

[31] Suhas Thejaswi, Bruno Ordozgoiti, and Aristides Gionis. "Diversity-Aware k-median: Clustering with Fair Center Representation". In: *Machine Learning and Knowledge Discovery in Databases. Research Track*. Springer International Publishing, 2021.

[32] Ali Vakilian and Mustafa Yalçıner. "Improved approximation algorithms for individually fair clustering". In: *arXiv preprint arXiv:2106.14043* (2021).