The thesis titled "Approximation Algorithms for Clustering and Submodular Facility Location Problems" submitted by Fateme Abbasi mainly considers three problems, and each result is presented in Chapters 2-4.

Chapter 2 of the thesis considers the Socially Fair (k, z)-Clustering problem. Given a set of points P, m groups $S_1, \ldots, S_m \subseteq P$, a set of potential centers F, a metric δ on $P \cup F$, weight function $w : P \to \mathbb{R}_+$, and a positive integer k, the problem aims at finding a k-element set of centers $X \subseteq F$ that minimizes $\max_{i \in [m]} \sum_{p \in S_i} w(p)\delta(p, X)^z$.

In the first part of the chapter, for the discrete d-dimensional Euclidean case, an FPT $(3^z(1-\eta))$-approximation algorithm for Socially Fair (k, z)-Clustering with the running time of $2^{O(k \log k)}\mathsf{poly}(m, n, d)$ is presented, where $\eta > 0$ is an absolute constant. This result is particularly interesting as it beats the approximability lower bound of 3^z for general metrics. The analysis involves a strengthened version of the projection lemma of Goyal and Jaiswal [73], giving a slightly better bound for those points that are far from the centers of a bi-criteria solution computed at the beginning of the algorithm.

The second result in this chapter is an FPT hardness result for the discrete Euclidean problem that rules out an EPAS. This result is based on a reduction from Multi-Colored Independent Set using Ta-Shma's construction [116] of η-balanced error-correcting codes. While the first result in this chapter serves to distinguish the discrete Euclidean case of the problem from the general case, this second result differentiates the discrete Euclidean case from the continuous one.

The above hardness holds for logarithmic dimension; as the last result in the chapter, the thesis gives an EPAS for doubling metric with sublogarithmic dimension. This algorithm constructs a coreset by considering a weighted re-formulation of the problem.

Chapter 3 presents an EPAS for the Norm k-Clustering problem. In this problem, the objective function is defined by a function $f : \mathbb{R}^P \to \mathbb{R}$ that maps $(\delta(p, X))_{p \in P}$ to a number, where X denotes the set of chosen centers (the rest of the notation is identical to the previous chapter). This function f is given as input and required to be a *norm* (see Chakrabarty and Swamy [29] for definition). This problem is a common generalization of various problems, and in fact, the thesis gives a lengthy list of such problems. This result is quite interesting as it gives a "unified" algorithm for these commonly studied problems.

The main technical tool is the (algorithmic) ϵ-scatter dimension. The ϵ-scatter dimension of a metric space is the maximum length of a sequence constructed by two "players," the *center player* and the *point player*. The center player always names a center that is close enough to cover the points named so far, while the point player tries to name a point that is far from the last center. The *algorithmic* ϵ-scatter dimension is similarly defined, but the centers are named by an algorithm called Ball Intersection, the point player specifies a radius too, and the number of triples per radius instead of the entire sequence length determines the dimension.

Based on this, the weighted k-median problem is first solved using an algorithm that guesses the optimal solution and then randomly determines a set of points that must be close to each open centers. These random sets are incrementally constructed and therefore will contain only a few points, limited by the algorithmic ϵ-scatter dimension. This bound also implies that the random choice will be valid with some probability. The thesis then extends this result to the general norm function by observing that the general problem can be considered as the intersection of

weighted k-median problems.

The chapter concludes by showing a multiple number of metrics that have bounded (algorithmic) ϵ-scatter dimension, proving that they can be used in conjuction with the algorithm presented.

Finally, Chapter 4 considers the Submodular Facility Location problem. This problem differs from the "standard" facility location problem in that the opening cost of a facility is given as a submodular function of its assigned clients. It is important to note that this submodular function is shared by all facilities.

The given algorithm begins by solving the natural configuration LP, where each variable corresponds to a "star" of clients centered around a facility. Then it samples integral (partial) assignments from the distribution defined by the solution. This sampling is repeated $O(\log \log N)$ times, where N is the number of facilities and clients, introducing a blow-up in the assignment costs but reducing opening costs.

In order to complete this partial solution into a feasible one, it first embeds the input metric into a tree metric. Note that this inflates the assignment costs but each client is already covered with high probability. Then by relocating the (mapped) clients to a lowest node whose subtree contains a large fraction of facilities covering that client, we can transform the given problem into a Descendent-Leaf Assignment (DLA) problem. The algorithm of Bosman and Olver [22] can then be adapted to solve this DLA problem, yielding the desired approximation ratio.

This result also extends to generalizations of the problem, named MULTSFL (and ADDSFL), where the opening costs are multiplied (and added, respectively) by facility-specific factors. The result also gives an algorithm for the Universal Stochastic Facility Location problem, which is a stochastic version of the regular facility location problem where clients are stochastically activated and we aim at optimizing the expected cost of the solution that deterministically assigns each client to a facility.

The thesis considers two interesting clustering problems, namely Socially Fair (k, z)-Clustering and Norm k-Clustering, and present new FPT approximation algorithms. The results for the former problem are interesting as they reveal new insights on how the problem's approximability changes depending on the structure of the metric cost. For the latter problem, the thesis gives a unified approach to a multiple number of previously studied problems. Both results open interesting future research directions as well, and the techniques used in the analyses appear to be of their own interest. The thesis continues to consider the Submodular Facility Location problem, and give a nice and clean approximation algorithm. This gives an asymptotically better approximation ratio compared to the previously known; a major open question is whether constant-factor approximation is possible.

Overall, the thesis presents intriguing and meaningful discoveries in the field of approximation algorithms, and also poses interesting future research directions. Taken together, these points indicate that the thesis is appropriate for the award of a doctoral degree and reflects a strong level of scholarly achievement.

I hereby submit this review.

Hyung-Chan An

Associate Professor, Yonsei University