Abstract

Dasgupta recently introduced a cost function for the hierarchical clustering of a set of points given pairwise similarities between them. He showed that this function is NP-hard to optimize, but a top-down recursive partitioning heuristic based on an α_n-approximation algorithm for uniform sparsest cut gives an approximation of $O(\alpha_n \log n)$ (the current best algorithm has $\alpha_n = O(\sqrt{\log n})$). We show that the aforementioned sparsest cut heuristic in fact obtains an $O(\alpha_n)$-approximation. The algorithm also applies to a generalized cost function studied by Dasgupta. Moreover, we obtain a strong inapproximability result, showing that the Hierarchical Clustering objective is hard to approximate to within any constant factor assuming the *Small-Set Expansion (SSE) Hypothesis*. Finally, we discuss approximation algorithms based on convex relaxations. We present a spreading metric SDP relaxation for the problem and show that it has integrality gap at most $O(\sqrt{\log n})$. The advantage of the SDP relative to the sparsest cut heuristic is that it provides an explicit lower bound on the optimal solution and could potentially yield an even better approximation for hierarchical clustering. In fact our analysis of this SDP served as the inspiration for our improved analysis of the sparsest cut heuristic. We also show that a spreading metric LP relaxation gives an $O(\log n)$-approximation.