The r-Gather Problem in Euclidean Space

Esther M. Arkin* Jie Gao† Matthew P. Johnson‡ Joseph S. B. Mitchell*
Jiemin Zeng†

1. INTRODUCTION

Given a set of n points $P = \{p_1, p_2, \ldots, p_n\}$ in Euclidean space and a value r, the aim of the r-gather problem is to cluster the points into groups of at least r points each such that the largest diameter of the clusters is minimized. We have two definitions of the diameter of a cluster: the distance between the furthest pair of points and the diameter of the smallest enclosing circle.

One motivation of this version of clustering is from location privacy in wireless networking. With the ubiquitous use of GPS receivers on mobile devices, it is now common practice that the locations of these mobile devices are recorded and collected. This raised privacy concerns as location information is sensitive and can be used to identify the user of the devices. One common practice in the treatment of these location data is to adopt the k-anonymity criterion [5], which says that the locations are grouped into clusters, each of at least k points. The cluster is used to replace individual locations such that any one user is not differentiated from $k-1$ others. Thus minimizing the diameter of the clusters can lead to location data with best accuracy while not intruding user privacy.

The r-gather problem is shown to be NP-hard to approximate at a ratio better than 2 when $r > 6$ and the points are in a general metric by Aggarwal et al.[1]. They also provide a 2-approximation algorithm. The approximation algorithm first guesses the optimal diameter and greedily selects clusters with twice the diameter. Then, a flow algorithm is constructed to assign at least r points to each cluster. This procedure is repeated until a good guess is found. Note that this solution only selects input points as cluster centers.

Armon [3] extended the result of Aggarwal et al. by proving it is NP-hard to approximate at a ratio better than 2 for the general metric case when $r > 2$. He also specifies a generalization of the r-gather clustering problem named the r-gathering problem which also considers a set of potential cluster centers (refered to as potential facility locations in Armon's paper) and their opening costs in the final optimization function. They provide a 3-approximation to the min-max r-gathering problem and prove it is NP-hard to have a better approximation factor. They also provide various approximation algorithms for the min-max r-gathering problem with the proximity requirement; a requirement for all points to be assigned to their nearest cluster center.

For the case where $r = 2$, both [2] and [4] provide polynomial time exact algorithms. Shalita and Zwick's [4] algorithm runs in $O(mn)$ time, for a graph with n nodes and m edges.

2. NEW RESULTS

For the application of protecting location privacy, the data points are actually in Euclidean spaces. Thus, we ask whether the hardness of approximation still holds in the Euclidean space. In the following, we assume that the input points are in the Euclidean plane.

For the case where the diameter of a cluster is the diameter of the smallest covering disk, we show it is NP-hard to approximate better than $\sqrt{13}/2 \approx 1.802$ when $r = 3$ and $\frac{\sqrt{35}+\sqrt{3}}{4} \approx 1.912$ when $r \geq 4$.

For the case where the diameter of a cluster is the distance between the furthest pair of points, then it is NP-hard to approximate better than $\sqrt{2+\sqrt{3}} \approx 1.931$ when $r = 3$ or 4 and 2 when $r \geq 5$.

3. HARDNESS PROOF

THEOREM 1. *The r-gather problem for the case where the diameter of a cluster is measured by the furthest distance between two points is NP-hard to approximate better than a factor of 2 when $r \geq 5$.*

PROOF. Our reduction is from the NP-hard problem, planar 3SAT. Given a formula in 3CNF composed of variables $x_i, i = 1, \ldots, n$ and their complements $\overline{x_i}$, we construct an instance of r-gather on the plane. Figure 1 illustrates a clause gadget of the clause $C = x_i \vee x_j \vee x_k$ and part of a variable gadget for x_i. In the figure, each point represents multiple points in the same location, the number of which is noted in parenthesis. All distances between groups of points connected by a line are distance 1 apart. Note that all clusters shown in the figure have a diameter of 1. If all clusters

*Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794. Partially supported by NSF (CCF-1526406). Email: esther.arkin@stonybrook.edu, jsbm@ams.stonybrook.edu.
†Department of Computer Science, Stony Brook University, Stony Brook, NY 11794. Email: {jgao, jiezeng}@cs.stonybrook.edu.
‡Department of Mathematics and Computer Science, Lehman College, Bronx, NY 10468, and PhD Program in Computer Science, CUNY Graduate Center, NY 10016. Email: mpjohnson@gmail.com.

have a diameter of 1, then we can signify the parity of a variable by whether solid or dashed clusters are chosen. Here the solid clusters signify a positive value for x_i that satisfies the clause since the center point of the clause gadget is successfully assigned to a cluster. Note that the variable gadget in Figure 1 swaps the parity of the signal sent away from the gadget. We also include a negation gadget shown in Figure 2 that swaps the parity of the signal and can be used when connecting parts of the variable gadget together. If an optimal solution to this r-gather construction can be found, the diameter of all clusters is 1.

The center point of the clause gadget must be assigned to a cluster that contains all r points of one of the variable clusters or else a cluster of diameter 2 is forced. WLOG, let the center point be clustered with the r points of the x_i gadget. What results is the solid clusters in Figure 1 are selected above the triangle splitter and the dashed clusters are selected below the splitter. The group of points at the top of the triangle splitter is unassigned to a cluster. It must merge with one of the neighboring clusters which results in a cluster of diameter 2. Therefore, it is NP-hard to approximate r-gather below a factor of 2 for $r \geq 5$. $\square$

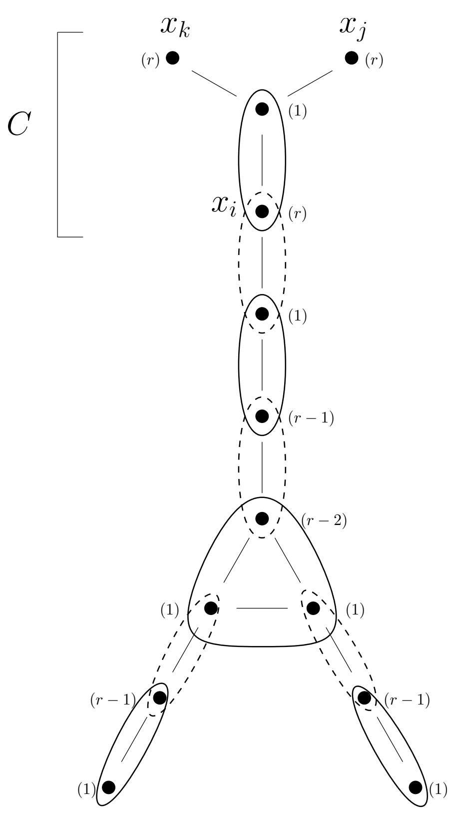

Figure 1: Clause and splitter gadget

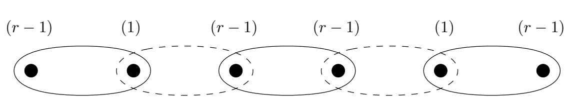

Figure 2: Signal negation gadget

THEOREM 2. *The r-gather problem for the case where the diameter of a cluster is measured by the diameter of the smallest covering disk is NP-hard to approximate better than a factor of* $\frac{\sqrt{35}+\sqrt{3}}{4} \approx 1.912$ *when* $r \geq 4$.

PROOF. The reduction is very simlar to the proof of Theorem 1. The only difference is the splitter which is illustrated in Figure 3. $\square$

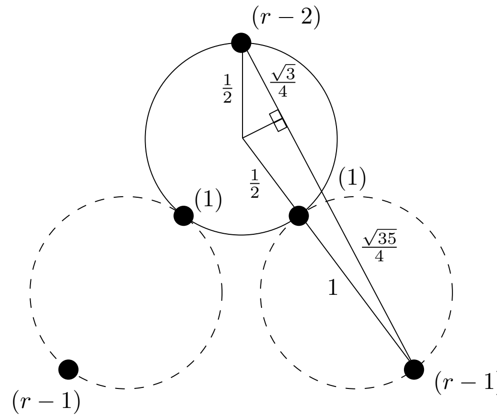

Figure 3: Splitter gadget

COROLLARY 1. *The r-gather problem in the L_1 and L_∞ metrics is NP-hard to approximate better than a factor of 2.*

THEOREM 3. *The r-gather problem for the case where the diameter of a cluster is the distance between the furthest pair of points, then it is NP-hard to approximate better than* $\sqrt{2+\sqrt{3}} \approx 1.931$ *when* $r = 3$ *or* 4.

THEOREM 4. *The r-gather problem for the case where the diameter of a cluster is measured by the diameter of the smallest covering disk is NP-hard to approximate better than a factor of* $\sqrt{13}/2 \approx 1.802$ *when* $r = 3$.

Corollary 1 is consequence of Theorem 1. Theorems 3 and 4 are proved with reductions from planar circuit SAT. The gadgets used in the reduction are similar to the splitter gadget used in the proof of Theorem 1. Details of the proofs are omitted due to space constraints.

4. REFERENCES

[1] G. Aggarwal, S. Khuller, and T. Feder. Achieving anonymity via clustering. In *In PODS*, pages 153–162, 2006.

[2] E. Anshelevich and A. Karagiozova. Terminal backup, 3D matching, and covering cubic graphs. *SIAM Journal on Computing*, 40(3):678–708, 2011.

[3] A. Armon. On min–max r-gatherings. *Theoretical Computer Science*, 412(7):573–582, 2011.

[4] A. Shalita and U. Zwick. Efficient algorithms for the 2-gathering problem. *ACM Transactions on Algorithms (TALG)*, 6(2):34, 2010.

[5] L. Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10:557–570, Oct. 2002.