Discrete Mathematics, Algorithms and Applications

r-Gatherings on a Star and Uncertain r-Gatherings on a Line

Shareef Ahmed

Graph Drawing and Information Visualization Laboratory,
Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
shareefahmed@cse.buet.ac.bd

Shin-ichi Nakano

Gunma University, Kiryu 376-8515, Japan
nakano@cs.gunma-u.ac.jp

Md. Saidur Rahman

Graph Drawing and Information Visualization Laboratory,
Department of Computer Science and Engineering,
Bangladesh University of Engineering and Technology, Dhaka, Bangladesh
saidurrahman@cse.buet.ac.bd

Received Day Month Year
Revised Day Month Year
Accepted Day Month Year
Published Day Month Year

Let C be a set of n customers and F be a set of m facilities. An r-gather clustering of C is a partition of the customers in clusters such that each cluster contains at least r customers. The r-gather clustering problem asks to find an r-gather clustering which minimizes the maximum distance between a pair of customers in a cluster. An r-gathering of C to F is an assignment of each customer $c \in C$ to a facility $f \in F$ such that each facility has zero or at least r customers. The r-gathering problem asks to find an r-gathering that minimizes the maximum distance between a customer and his/her facility. In this work we consider the r-gather clustering and r-gathering problems when the customers and the facilities are lying on a "star". We show that the r-gather clustering problem and the r-gathering problem with customers and facilities on a star with d rays can be solved in $O(n + d^d r^d dr \log d)$ and $O(n + m + (d + \log m)d^4 r^2 + d^d r^d 2^d dr \log d)$ time, respectively. Furthermore, we prove the hardness of a variant of the r-gathering problem, called the min-max-sum r-gathering problem, even when the customers and the facilities are on a star. We also study the r-gathering problem when the customers and the facilities are on a line, and each customer location is uncertain. We show that the r-gathering problem can be solved in $O(nk+mn\log n+(m+n\log kn+nr^{\frac{n}{r}})\log mn)$ and $O(mn\log n+(n\log n+m)\log mn)$ time when the customers and the facilities are on a line, and the customer locations are given by piecewise uniform functions of at most

$k+1$ pieces and "well-separated" uniform distribution functions, respectively.

Keywords: r-Gathering problem; Facility location problem; Clustering.

Mathematics Subject Classification: 11xxx, 11xxx, 11xxx

1. Introduction

The facility location problem and many of its variants are well studied [9]. In this paper we study some variants of the facility location problem.

Let C be a set of n customers and $d(p, q)$ be the distance between $p, q \in C$. An r-gather clustering R of C is a partition of the customers of C in clusters such that each cluster contains at least r customers. The cost $cost(\mathcal{C})$ of a cluster $\mathcal{C}$ is the maximum distance between a pair of customers in $\mathcal{C}$. The cost $cost(R)$ of an r-gather clustering R is the maximum cost among the costs of the clusters. The r-gather clustering problem asks to find an r-gather clustering of C with minimum cost [4], and such a clustering is called an optimal r-gather clustering.

Let C be a set of n customers and F be a set of m facilities, $d(c, f)$ be the distance between $c \in C$ and $f \in F$. An r-gathering of C to F is an assignment $A : C \to F$ such that each facility has zero or at least r customers assigned to it. The cost of an r-gathering is $\max_{c \in C}\{d(c, A(c))\}$, which is the maximum distance between a customer and his/her facility. The r-gathering problem asks to find an r-gathering of C to F having the minimum cost [8]. This problem is also known as the min-max r-gathering problem. The other version of the problem is known as the min-sum r-gathering problem that asks to find an r-gathering which minimizes $\sum_{c \in C} d(c, A(c))$ [14,11]. In this paper we consider the min-max r-gathering problem and we use the term r-gathering problem to refer the min-max version unless specified otherwise.

Assume we wish to set up emergency shelters for residents C living on a locality so that each shelter can accommodate at least r residents. We also wish to locate the shelters so that evacuation time span can be minimized. We can model this scenario by the r-gather clustering problem. For each cluster in an optimal r-gather clustering of C, we set up a shelter at the center of the minimum enclosing circle covering the residents of the cluster, and assign the residents to the shelter. If a set F of possible locations for the shelters is also given, then the scenario can be modeled by the r-gathering problem. In this case, an r-gathering corresponds to an assignment of residents to shelters so that each "open" shelter serves at least r residents and the r-gathering problem finds the r-gathering minimizing the evacuation time span.

Both r-gather clustering and r-gathering problems are NP-complete in general [4,8]. For the r-gather clustering problem, a 2-approximation algorithm is known [4]. For the r-gathering problem, a 3-approximation algorithm is known and it is proved that the problem cannot be approximated within a factor less than 3 for $r > 3$ unless $P = NP$ [8]. Recently, both problems are considered in a setting where all customers and facilities are lying on a line. An $O(n \log n)$ time algorithm [7], and an $O(rn)$ time algorithm [15], and an $O(n)$ time algorithm [16] are known for the r-gather

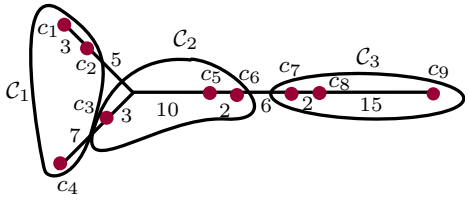


Fig. 1. An optimal 3-gather clustering on a star.

clustering problem when all the customers are on a line. For the r-gathering problem an $O((n+m)\log(n+m))$ time algorithm [7], an $O(n+m\log^2 r+m\log m)$ time algorithm [12], an $O(n+r^2m)$ time algorithm [15], and an $O(n+m)$ time algorithm [16] are known when all the customers and facilities are on a line. Recently, the r-gather clustering problem is studied on mobile setting and a 4-approximation distributed algorithm is known [22].

In this paper, we first consider both the r-gathering clustering and r-gathering problem when the customers are on a star. When the customers are on a line, each cluster of an optimal r-gather clustering consists of consecutive customers on the line [15]. However, when the customers are on a star, some clusters may not consists of consecutive customers in the optimal r-gather clustering. For example, see Figure 1. We can observe that at least one cluster consists of non-consecutive customers in any optimal r-gather clustering. Figure 1 demonstrates an optimal r-gather clustering for this scenario.

In this paper we give an $O(n+d^dr^ddr\log d)$ time algorithm for r-gather clustering problem on a star, and an $O(n+m+(d+\log m)d^4r^2+d^dr^d2^ddr\log d)$ time algorithm for the r-gathering problem on a star, where d is the number of rays that form the star. We also proved the hardness of a variant of the r-gathering problem, called the min-max-sum r-gathering problem.

We also consider the r-gathering problem when the customer and the facilities are on a line, and each customer location is uncertain. Let $F=\{f_1,f_2,\cdots,f_m\}$ be a set of m facilities, and $C=\{c_1,c_2,\cdots,c_n\}$ be a set of n customers where each customer location is a random variable (Although random variables are traditionally denoted by capital letters, we denote them by small letters for consistency). The *uncertain r-gathering problem* asks to find an r-gathering such that the maximum expected distance between a customer and his/her facility is minimum. Note that, the uncertain r-gathering problem is NP-hard, since it contains the deterministic version as a special case.

Problems under uncertain settings has become much popular recently. Uncertainty in data occurs because of noise in measured data, sampling inaccuracy, limitation of resources, etc. Hence, uncertainty is ubiquitous in practice and managing the uncertain data has gained much attention [1,2,3,18]. Some variants of the facility location problem have also been investigated under some uncertain settings. Setting

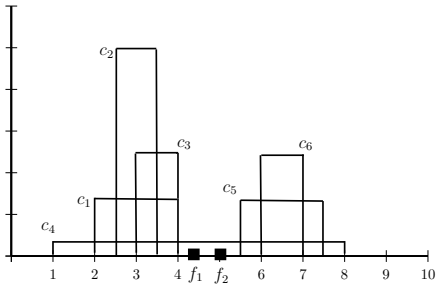


Fig. 2. An instance of uncertain 3-gathering problem.

up a facility is costly and each facility is supposed to serve for a long period of time. On the other hand existence, location and demand of a client can change over time. Thus it is important to set up facilities by keeping the uncertainty in mind. For the detailed state of the art of uncertain facility location problem, we refer to the survey of Snyder [17]. There are two models for uncertainty: one is existential model [13,21] and the other is locational model [1,2,19]. In the existential model, the existence of each customer is uncertain. Thus each customer has a specific location and there is a probability for the existence of each customer. In the locational model each customer is certain to exist, but his/her position is uncertain and defined by a probability density function. In this paper we consider the locational model of uncertainty. For customer locations, we consider two probability density functions: piecewise uniform function (histogram) and "well-separated" uniform distribution function.

When the customer and facility locations are deterministic and on a line, there is an optimal r-gathering where the customers assigned to each facility are consecutive on the line [15]. However, when the customer locations are uncertain, finding a suitable ordering of the customers is difficult. For example, consider Figure 2. Here each customer location has a uniform distribution. The instance has only one optimal r-gathering, and in the optimal r-gathering c_1, c_2, c_4 are assigned to f_1, and c_3, c_4, c_5 are assigned to f_2. Although midpoint (or mean) of c_3 is on the left of the midpoint (or mean) of c_4, in the optimal solution c_4 is assigned to f_1 and c_3 is assigned to f_2. We show that the r-gathering problem can be solved in $O(nk+mn\log n+(m+n\log kn+nr^{\frac{n}{r}})\log mn)$ and $O(mn\log n+(n\log n+m)\log mn)$ time when the customers and the facilities are on a line, and the customer locations are given by piecewise uniform functions of at most $k+1$ pieces and "well-separated" uniform distribution functions, respectively.

The rest of the paper is organized as follows. In Section 2 we define terms used in the paper. In Section 3 we give an algorithm for the r-gather clustering problem on a star. In Section 4 we give an algorithm for the r-gathering problem on a star. We show the hardness of the min-max-sum r-gathering problem on a star in Section 5. In Section 6, we give algorithms for uncertain r-gathering problem when customer

locations are specified by piecewise uniform functions and "well-separated" uniform distribution functions. Finally Section 7 is a conclusion. Preliminary versions of some results of this paper were presented at [6] and [5].

2. Preliminaries

In this section we define some terms used in this paper.

Let $\mathcal{L} = \{l_1, l_2, \cdots, l_d\}$ be a set of d rays where all the rays of $\mathcal{L}$ share a common source point o. We call the set $\mathcal{L}$ of rays a *star* and the common source point o the *center* of the star. The *degree* of a star is the number d of rays which form the star. The Euclidean distance between two points u, v on a ray is denoted by $d_E(u, v)$. The distance $d(p, q)$ between two points p and q in $\mathcal{L}$ is defined as $d(p, q) = d_E(p, q)$, if both p and q are on the same ray, and $d(p, q) = d_E(p, o) + d_E(o, q)$ otherwise. For ease of notation, the point where a customer c_i (respectively, a facility f_j) is located is denoted by c_i (respectively, f_j). A cluster consisting of customers from two or more rays is a *multi-ray cluster*, otherwise a *single-ray cluster*. Two customers p and q are the *end-customers* of a cluster $\mathcal{C}$ if $d(p, q) = cost(\mathcal{C})$.

In an r-gather clustering, the cost of a cluster $\mathcal{C}$, denoted by $cost(\mathcal{C})$, is defined as $\max_{p,q\in\mathcal{C}} d(p, q)$. The following result is known [15] regarding the r-gather clustering problem. Note that any cluster with $2r$ or more customers can be divided into some clusters so that each of which has at most $2r-1$ customers and at least r customers.

Lemma 2.1 ([15]). *There is an optimal r-gather clustering in which each cluster has at most* $2r - 1$ *customers.*

In the r-gathering problem, a facility with one or more customers is called an *open facility*. $A(c)$ denotes the facility to which a customer c is assigned in an assignment A. The cost of a facility f, denoted by $cost(f)$, is $\max\{d(f, c_i)|A(c_i) = f\}$ if f has one or more customers, and is 0 if f has no customer.

When the customer locations are on a line and uncertain, each uncertain customer c_i is associated with a probability density function (PDF), denoted by $g_i(x)$ where x is a point on the line. The expected distance between a facility f_j and an uncertain customer c_i, denoted by $E[d(c_i, f_j)]$, is $\int_{-\infty}^{\infty} d(x, f_j)g_i(x)dx$.

3. *r*-Gather Clustering on a Star

In this section we give an algorithm for the r-gather clustering problem on a star. Let C be a set of customers on a star $\mathcal{L} = \{l_1, l_2, \cdots, l_d\}$ of d rays with center o. We have the following lemma.

Lemma 3.1. *There is an optimal r-gather clustering such that, for each ray* l_i*, the customers on* l_i *assigned to the multi-ray clusters are consecutive customers on* l_i *including the nearest customer to o on* l_i*.*

Proof. A pair of customers c_m, c_s on l_i is called a reverse pair if c_m is assigned to a multi-ray cluster, c_s is assigned to a single-ray cluster, and $d(o, c_s) < d(o, c_m)$.

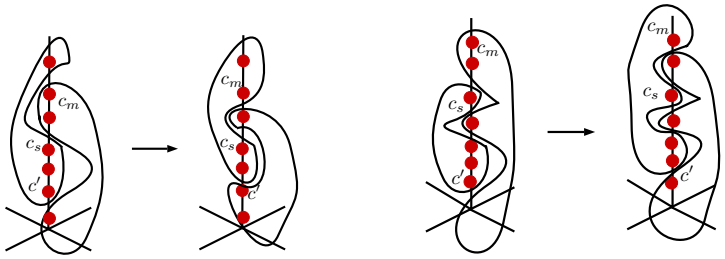


Fig. 3. (a) Illustration of Case 1 and (b) illustration of Case 2 of proof of Lemma 3.1.

Assume for a contradiction that R is an optimal r-gather clustering with the minimum number of reverse pairs but the number is not zero. Let c_s and c_m be a reverse pair on l_i with maximum $d(o, c_m)$. Let $\mathcal{C}_s$ and $\mathcal{C}_m$ be the clusters containing c_s and c_m, respectively. We have two cases.

Case 1: $\mathcal{C}_s$ has a customer c on l_i with $d(o, c_m) < d(o, c)$.
Let c' be the nearest customer to o in $\mathcal{C}_s$. Replacing $\mathcal{C}_s$ and $\mathcal{C}_m$ in the clustering by $\mathcal{C}_s \setminus \{c'\} \cup \{c_m\}$ and $\mathcal{C}_m \setminus \{c_m\} \cup \{c'\}$ generates a new r-gather clustering with less reverse pairs as illustrated in Figure 3(a). A contradiction. Note that $cost(\mathcal{C}_s \setminus \{c'\} \cup \{c_m\}) \leq cost(\mathcal{C}_s)$ and $cost(\mathcal{C}_m \setminus \{c_m\} \cup \{c'\}) \leq cost(\mathcal{C}_m)$ hold.
Case 2: Otherwise. (Thus $d(o, c) < d(o, c_m)$ for every customer c in $\mathcal{C}_s$.)
The same replacing results in a new r-gather clustering with less reverse pairs as illustrated in Figure 3(b). A contradiction. Note that $cost(\mathcal{C}_s \setminus \{c'\} \cup \{c_m\}) \leq cost(\mathcal{C}_m)$ and $cost(\mathcal{C}_m \setminus \{c_m\} \cup \{c'\}) \leq cost(\mathcal{C}_m)$ hold. □

Lemma 3.2. *If an optimal r-gather clustering has multi-ray clusters, then there is an optimal r-gather clustering where at most one multi-ray cluster contains more than r customers.*

Proof. Assume for a contradiction that there is no optimal r-gather clustering where at most one multi-ray cluster contains more than r customers. Let R be an r-gather clustering with the minimum number of multi-ray clusters having more than r customers. Let $\mathcal{C}_i$ and $\mathcal{C}_j$ be two multi-ray clusters having more than r customers. Let p_i, q_i be the two end-customers of $\mathcal{C}_i$ and p_j, q_j be the two end-customers of $\mathcal{C}_j$. Without loss of generality, assume that q_j is the closest customer to o among the four end-customers. Let $\mathcal{C}'_j \subset \mathcal{C}_j$ be $\{c \in \mathcal{C}_j | d(o, c) > d(o, q_j)\}$. Any customer $c \in \mathcal{C}'_j$ must be on the same ray as p_j, otherwise q_j would not be an end-customer of $\mathcal{C}_j$. We have two cases.
Case 1: $|\mathcal{C}'_j| < r$.
Let $\mathcal{C}''_j$ be a set of $|\mathcal{C}_j| - r$ arbitrary customers from $\mathcal{C}_j \setminus \mathcal{C}'_j$. We now derive a new r-gather clustering R' by replacing $\mathcal{C}_i$ and $\mathcal{C}_j$ by $\mathcal{C}_i \cup \mathcal{C}''_j$ and $\mathcal{C}_j \setminus \mathcal{C}''_j$. Since q_j is the closest customer to o among the four end-customers p_i, q_i, p_j, q_j and $d(o, c) \leq$

$d(o, q_j)$ for any customer $c \in \mathcal{C}''_j$, we have $d(o, c) \leq d(o, p_i)$ and $d(o, c) \leq d(o, q_i)$. Thus the cost of $\mathcal{C}_i \cup \mathcal{C}''_j$ does not exceed the cost of $\mathcal{C}_i$. Hence the cost of R' is not greater than the cost of R. Thus R' has less multi-ray clusters with more than r customers, a contradiction.

Case 2: Otherwise. Thus $|\mathcal{C}'_j| \geq r$.

In this case we derive a new r-gather clustering R' by replacing $\mathcal{C}_i$ and $\mathcal{C}_j$ by $\mathcal{C}_i \cup (\mathcal{C}_j \setminus \mathcal{C}'_j)$ and $\mathcal{C}'_j$. In this case, $\mathcal{C}'_j$ is a single-ray cluster. By a similar argument of Case 1, the cost of R' does not exceed the cost of R. Thus R' has less multi-ray clusters with more than r customers than R, a contradiction. □

We solve the problem by computing all possible r-gather clusterings consisting of only multi-ray clusters of suitable $S \subset C$ near o and an r-gathering clustering for each remaining ray. We now give the following lemma, which is used in the proof of Lemma 3.4 and Lemma 3.6.

Lemma 3.3. *If there is an optimal r-gather clustering consisting of only multi-ray clusters, then there is an optimal r-gather clustering with the multi-ray cluster consisting of the farthest customer from o and his/her $r-1$ nearest customers.*

Proof. Let p be the farthest customer from o and let N be the $r-1$ nearest customers of p. Assume for a contradiction that there is no optimal solution in which $N \cup \{p\}$ is a cluster. We first prove that $N \cup \{p\}$ is contained in the same cluster. Let R be an optimal solution with cluster $\mathcal{C}_p$ containing p has the maximum number of customers in N. Let q be a customer in N assigned to a cluster $\mathcal{C}_q \neq \mathcal{C}_p$. Since the number of customers in $\mathcal{C}_p$ is at least r, there is a customer $p' \in \mathcal{C}_p$ not in N. Let q' be the farthest customer from o in $\mathcal{C}_q \setminus \{q\}$. We now derive a new r-gather clustering by replacing $\mathcal{C}_p$ and $\mathcal{C}_q$ by $\mathcal{C}_p \setminus \{p'\} \cup \{q\}$ and $\mathcal{C}_q \setminus \{q\} \cup \{p'\}$. Thus a contradiction. Note that, $cost(\mathcal{C}_p \setminus \{p'\} \cup \{q\}) \leq cost(\mathcal{C}_p)$ and $cost(\mathcal{C}_q \setminus \{q\} \cup \{p'\}) \leq \max\{cost(\mathcal{C}_p), cost(\mathcal{C}_q)\}$, since $d(o, p) \geq d(o, q')$.

We now prove that $N \cup \{p\}$ form a multi-ray cluster. Assume for a contradiction that there is no optimal r-gather clustering where $N \cup \{p\}$ is a cluster. Let R' be an optimal r-gather clustering with cluster $\mathcal{C}_p$ containing p having the minimum number of customers not in N. Let p'' be the farthest customer in $\mathcal{C}_p$ not in the ray l_p containing p, and $\mathcal{C}_s$ be a cluster in R' other than $\mathcal{C}_p$. Let s be the farthest customer from o in $\mathcal{C}_s$. We now derive a new r-gather clustering by replacing $\mathcal{C}_p$ and $\mathcal{C}_s$ with $\mathcal{C}_p \setminus \{p''\}$ and $\mathcal{C}_s \cup \{p''\}$ without increasing cost, a contradiction. Since $d(o, s) \leq d(o, p)$, we have $d(s, p'') \leq d(p, p'')$ and thus $cost(\mathcal{C}_s \cup \{p''\}) \leq cost(\mathcal{C}_p)$. □

We now have the following lemma.

Lemma 3.4. *If an optimal r-gather clustering consists of only multi-ray clusters, then there is an optimal r-gather clustering with at most $d-1$ multi-ray clusters.*

Proof. We give a proof by induction on the number d of rays in the star. Clearly, the claim holds for $d = 2$, since in such case only one multi-ray cluster can exist.

Assume that the claim holds for any star with less than d rays. We now prove that the claim also holds for any star of d rays. Assume for a contradiction that there is no optimal solution with at most $d-1$ multi-ray clusters. Let p be the farthest customer from o. By Lemma 3.3, there is an optimal r-gather clustering with the multi-ray cluster $\mathcal{C}_p$ containing p and his/her $r-1$ nearest customers, denoted by N. Let l_p be the ray containing p. Since $\mathcal{C}_p$ is a multi-ray cluster and N consists of the $r-1$ nearest neighbors of p, all the customers on l_p are contained in $\mathcal{C}_p$. Thus the customers in $C \setminus \mathcal{C}_p$ are lying on other $d-1$ rays except l_p. By inductive hypothesis there is an optimal r-gather clustering of $C \setminus \mathcal{C}_p$ with at most $d-2$ multi-ray clusters. Thus the claim holds. ◻

Algorithm 1 Multi-rayClusters(C, d)

Input A set C of customers on a star, and degree d of star
Output An optimal r-gather clustering with only multi-ray clusters if exists

if $|C| < r$ or the number of rays containing at least one customer is at most one **then**
  return $\emptyset$
end if
$i \leftarrow 1$
while $|C| \neq 0$ **do**
  if $|C| < 2r$ **then**
    Create new cluster $\mathcal{C}_i = C$
  else
    $p \leftarrow$ farthest customer from o in C
    $\mathcal{C}_i \leftarrow \{p, p_1, p_2, \cdots, p_{r-1}\}$ where p_j is the j-th nearest customer of p in C
  end if
  if $\mathcal{C}_i$ is a single-ray cluster **then**
    return $\emptyset$
  end if
  $C \leftarrow C \setminus \mathcal{C}_i$
  $i \leftarrow i + 1$
end while
return $\{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \cdots, \mathcal{C}_{i-1}\}$

Corollary 3.5. *If an optimal r-gather clustering consists of only multi-ray clusters, then C has at most $(d-2)r + 2r - 1 = dr - 1$ customers.*

We now give an outline of our algorithm which constructs an optimal r-gathering clustering on a star. We first choose set S of customers for multi-ray clusters. For each possible set S of customers we find the optimal r-gather clustering consisting

of only multi-ray clusters, by repeatedly searching for the farthest customer from o and his/her $r-1$ nearest customer as a multi-ray cluster of the remaining set of customers, by the algorithm **Multi-rayClusters**.

We now have the following lemma.

Lemma 3.6. *Let $R = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3, \cdots\}$ be the clusters computed by Algorithm **Multi-rayClusters**. If R has only multi-ray clusters, then R is an optimal r-gather clustering of C.*

Proof. The proof of this lemma is immediate from Lemma 3.3. □

Lemma 3.7. *Algorithm **Multi-rayClusters** runs in $O(dr \log d)$ time.*

Proof. To construct each cluster, Algorithm **Multi-rayClusters** first picks the farthest customer p in C, which can be done in $O(d)$ time. Each of the remaining customers in the cluster is either on the same ray as p or on a different ray. Each customer on the same ray as p can be determined in constant time. Each customer on different rays can be determined in $O(\log d)$ time by maintaining a min-heap of the closest customer to o on each remaining rays. Thus, each cluster can be formed in $O(r \log d)$ time, and constructing all clusters takes $O(dr \log d)$ time. □

We now give an algorithm ***r*GatherClusteringOnStar** to construct an optimal r-gather clustering of C on a star. We have the following theorem.

Algorithm 2 rGatherClusteringOnStar(C, d)

Input A set C of customers on a star, and degree d of star
Output An optimal r-gather clustering

if $|C| < r$ **then**
  return $\emptyset$
end if
$Best \leftarrow \emptyset$
for each set S consists of at most $dr-1$ customers near to o **do**
  $R_m \leftarrow$ Multi-rayClusters(S, d)
  $R_i \leftarrow r$-gather clustering of customers of C lying on l_i by [16]
  $R \leftarrow R_m \cup R_1 \cup R_2 \cup \cdots \cup R_d$
  if R is better than $Best$ **then**
    $Best \leftarrow R$
  end if
end for
return $Best$

Theorem 3.8. *The algorithm* ***r GatherClusteringOnStar*** *constructs an optimal r-gather clustering of C on star in $O(n + d^d r^d dr \log d)$ time.*

Proof. We first prove that the algorithm rGatherClusteringOnStar(C, d) correctly produces an optimal r-gather clustering. By Lemma 3.1 multi-ray clusters in an optimal r-gathering are located near o, and by Corollary 3.5 the number of customers in the multi-ray clusters is at most $dr-1$. The algorithm ***r*GatherClusteringOnStar** checks every possible partition of C into (1) S consisting of customers for multi-ray clusters (near o), and (2) the remaining customers for single-ray clusters (far from o), and finds r-gather clusterings separately, then obtains an r-gather clustering of C by combining them. Finally outputs the best one.

We now estimate the running time of the algorithm. Let n_i be the number of customers on set S lying on ray l_i. By construction of set S, $\sum_{i=1}^{d} n_i \leq dr-1$ holds because possible S consists of at most $dr-1$ consecutive customers in each ray close to o. The number of ways $\sum_{i=1}^{d} n_i$ equals to some $x < dr$ is $\binom{x+d-1}{d-1}$, which is $O\left((dr)^{d-1}\right)$. Therefore, the number of ways S can be formed is $O(d^d r^d)$. For each such S we compute an r-gather clustering consists of only multi-ray clusters by algorithm Multi-rayClusters which runs in $O(dr \log d)$ time. We also compute single-ray clusters for the remaining customers (far from o). Rather than computing the single-ray clusters each time in the loop, we compute the r-gather clustering for customers consisting of i farthest customers from o, for each i, and for each ray in $O(n)$ time in total [16]. Thus, to compute all the required cases for single-ray cluster we need total $O(n)$ time. Therefore, the time complexity of the algorithm is $O(n + d^d r^d dr \log d)$. ◻

If d is constant then the running time of the algorithm **rGatherClusteringOnStar** is polynomial.

4. *r*-Gathering on a Star

In this section we give an algorithm for the r-gathering problem on a star.

Let C be a set of customers and F be a set of facilities on a star $\mathcal{L} = \{l_1, l_2, \cdots, l_d\}$ of d rays with center o. In any optimal r-gathering each open facility serves at least r customers. However the number of customers assigned to an open facility can be more than $2r-1$. In such case we regard the set of customers assigned to a facility as the union of clusters $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_k$ sharing a facility and each of which satisfies $r \leq |\mathcal{C}_i| < 2r$. Thus we can think of the r-gathering problem in a similar way to the r-gather clustering problem in Section 3, and Lemma 2.1 holds for the clusters of any r-gathering. We denote by $A(\mathcal{C})$ the facility to which the customers in $\mathcal{C}$ are assigned in r-gathering A. We define the cost of a cluster $\mathcal{C}$, denoted by $cost(\mathcal{C})$, in r-gathering A as $\max_{c \in \mathcal{C}}\{d(c, A(c))\}$. It is easy to observe that Lemma 3.1 also holds for the clusters of r-gatherings. We now prove that Lemma 3.2 also holds for the r-gathering problem.

Lemma 4.1. *There is an optimal r-gathering where at most one multi-ray cluster has more than r customers.*

Proof. Assume for a contradiction that there is no optimal r-gathering having at most one multi-ray cluster of size more than r. Let A be an r-gathering with the minimum number of multi-ray clusters having more than r customers. Let $\mathcal{C}_i$ and $\mathcal{C}_j$ be two multi-ray clusters with more than r customers. Let $f_i = A(\mathcal{C}_i)$ and $f_j = A(\mathcal{C}_j)$. Let f_i and f_j are located on ray l_i and l_j, respectively. Without loss of generality, assume that $d(o, f_i) \leq d(o, f_j)$. Let $\mathcal{C}'_j$ be the subset of $\mathcal{C}_j$ located on l_j. We have two cases.
Case 1: $|\mathcal{C}'_j| < r$. Let $\mathcal{C}''_j$ be a set of $|\mathcal{C}_j| - r$ arbitrary customers from $\mathcal{C}_j \setminus \mathcal{C}'_j$. We now derive a new r-gathering A' by replacing $\mathcal{C}_i$ and $\mathcal{C}_j$ by $\mathcal{C}_i \cup \mathcal{C}''_j$ and $\mathcal{C}_j \setminus \mathcal{C}''_j$, and assigning $\mathcal{C}_i \cup \mathcal{C}''_j$ and $\mathcal{C}_j \setminus \mathcal{C}''_j$ to f_i and f_j, respectively. Note that $\mathcal{C}_j \setminus \mathcal{C}''_j$ has exactly r customers. Let c be a customer in $\mathcal{C}''_j$. Since f_i is closer to o than f_j and c is not on l_j, we have $d(f_i, c) \leq d(o, f_i) + d(o, c) \leq d(o, f_j) + d(o, c) = d(c, f_j)$. Thus the cost of $\mathcal{C}_i \cup \mathcal{C}''_j$ does not exceed the cost of $\max\{cost(\mathcal{C}_i), cost(\mathcal{C}_j)\}$. Hence the cost of A' is not greater than the cost of A. Thus A' has less multi-ray clusters with more than r customers, a contradiction.
Case 2: Otherwise. Thus $|\mathcal{C}'_j| \geq r$. In this case we derive a new r-gathering A' by replacing $\mathcal{C}_i$ and $\mathcal{C}_j$ by $\mathcal{C}_i \cup (\mathcal{C}_j \setminus \mathcal{C}'_j)$ and $\mathcal{C}'_j$, and assigning $\mathcal{C}_i \cup (\mathcal{C}_j \setminus \mathcal{C}'_j)$ and $\mathcal{C}'_j$ to f_i and f_j, respectively. In this case, $\mathcal{C}'_j$ is a single-ray cluster. By a similar argument of Case 1, the cost of A' does not exceed the cost of A. Thus A' has less multi-ray clusters having more than r customers than A, a contradiction. □

A customer on a ray $l \in \mathcal{L}$ is the *boundary customer* of l if it is the farthest customer on l from o. We now give the following lemma.

Lemma 4.2. *If there is an optimal r-gathering A with only multi-ray clusters, then there is an optimal r-gathering with a multi-ray cluster consisting of a boundary customer and his/her* $r-1$ *nearest neighbors.*

Proof. Let f be the farthest open facility from o in A and l be the ray containing f. We have two cases to consider.
Case 1: l **has a customer.** Let p be the boundary customer on l and N be the set of the $r-1$ nearest customers of p. We first prove that there is an optimal solution with the customers in $N \cup \{p\}$ are assigned to f. Assume for a contradiction that there is no optimal solution where $N \cup \{p\}$ is assigned to f. Let A be an optimal solution with the maximum number of customers in $N \cup \{p\}$ are assigned to f. Let $\mathcal{C}_p$ be the multi-ray cluster assigned to f, and q be a customer in $N \cup \{p\}$ but $q \notin \mathcal{C}_p$. Let q is assigned to f'. Since $\mathcal{C}_p$ has at least r customers, there is a customer $p' \in \mathcal{C}_p$ not in $N \cup \{p\}$ and lying on a ray except l. We now derive a new r-gathering A' by reassigning q to f and p' to f'. Since $d(o, f') \leq d(o, f)$, we have $d(f', p') \leq d(o, f') + d(o, p') \leq d(o, f) + d(o, p') = d(f, p')$. Now if q is (1) not on l or (2) q is on l with $d(o, q) \leq d(o, f)$ then $d(f, q) \leq d(f, p')$. Otherwise, q is on l

with $d(o,q) > d(o,f)$ holds, then we have $d(f,q) \leq d(f,p)$. Thus the cost of A' does not exceed the cost of A, and A' has more customers in $N \cup \{p\}$ assigned to f. A contradiction. Thus the customers in $N \cup \{p\}$ are contained in $\mathcal{C}_p$.

We now prove that $N \cup \{p\}$ form a multi-ray cluster. Assume for a contradiction that there is no optimal r-gathering in which $N \cup \{p\}$ is a cluster. Let A' be an optimal r-gathering with the cluster $\mathcal{C}_p$ containing p having the minimum number of customers not in $N \cup \{p\}$. Since $\mathcal{C}_p$ is a multi-ray cluster, $\mathcal{C}_p$ has a customer p' not in $N \cup \{p\}$ and lying on a ray except l. Let $\mathcal{C}_s$ be a cluster in A' other than $\mathcal{C}_p$ and $A'(\mathcal{C}_s) = f'$. We now derive a new r-gathering by replacing $\mathcal{C}_p$ and $\mathcal{C}_s$ by $\mathcal{C}_p \setminus \{p'\}$ and $\mathcal{C}_s \cup \{p'\}$. We now derive a new r-gathering by reassigning p' to f'. Since $d(o,f') \leq d(o,f)$, $d(p',f')$ does not exceed $d(p',f)$. A contradiction.

Case 2: l has no customer. We first prove that all customers are assigned to f. Assume for a contradiction that there is an open facility $f' \neq f$ to which some customers are assigned. Since f is the farthest open facility from o and there is no customer on l, we can reassign all customers to f' without increasing the cost of the r-gathering. A contradiction. Let p be a boundary customer on ray $l' \neq l$, and N be his/her $r-1$ nearest neighbors. Since $|C| \geq 2r$, we can form a cluster $\mathcal{C}_p$ with $N \cup \{p\}$. □

We now prove that Lemma 3.4 also holds for r-gathering.

Lemma 4.3. *If an optimal r-gathering consists of only multi-ray clusters, then there is an optimal r-gathering consisting of at most $d-1$ multi-ray clusters, where d is the number of rays containing a customer.*

Proof. We give a proof by induction on d.

We first show the claim holds for $d = 2$. Assume for a contradiction that there is no optimal r-gathering having at most one multi-ray cluster. Let A be an optimal r-gathering with the minimum number of multi-ray clusters, and f be the farthest open facility from o in A and l be the ray containing f. If there is no customer on l, then by Lemma 4.2(b) all customers are assigned to f and the farthest boundary customer p and his/her $r-1$ nearest customers N form a cluster $\mathcal{C}$. Otherwise by Lemma 4.2(a) the boundary customer p of l and his/her $r-1$ nearest customers N form a cluster $\mathcal{C}$. In both case either $\mathcal{C}$ or the other cluster is a single-ray cluster, a contradiction.

Now we consider for $d > 2$. Assume that the claim holds if the customers are on less than d rays. We now prove that the claim also holds if the customers are on exactly d rays. Assume for a contradiction that there is no optimal r-gathering having at most $d-1$ multi-ray clusters. Let A be an optimal r-gathering with the minimum number of multi-ray clusters. Let f be the farthest open facility from o in A. Let l be the ray containing f. We have the following two cases.

Case 1: There is a customer on l. Let p be the boundary customer of l and N be the $r-1$ nearest customers of p. By Lemma 4.2(a), there is an optimal r-gathering

consisting of only multi-ray clusters with cluster $\mathcal{C}_p = N \cup \{p\}$. Now the customers in $C \setminus \mathcal{C}_p$ are lying on other $d-1$ rays except l. By inductive hypothesis there is an optimal r-gathering of $C \setminus \mathcal{C}_p$ with at most $d-2$ multi-ray clusters. Thus the claim holds.

Case 2: Otherwise. By Lemma 4.2(b), there is an optimal r-gathering where all customers are assigned to f. Since there are at least d multi-ray clusters, the number of customers is at least dr. Thus there is a ray l' with r or more customers. We can form a single-ray cluster with the r customers on l. A contradiction. □

We now give algorithm Multi-rayClusters2. If there is an optimal r-gathering with only multi-ray clusters, then the algorithm finds such an r-gathering, by repeatedly removing a cluster ensured by Lemma 4.2.

Lemma 4.4. *If there is an optimal r-gathering consisting of only multi-ray clusters, then Algorithm **Multi-rayClusters2** finds an optimal r-gathering. The running time of the algorithm is $O((d + \log m)d^4r^2 + 2^d dr \log d)$.*

Proof. If there is an optimal r-gathering with only multi-ray clusters, then, by repeatedly removing a cluster ensured by Lemma 4.2, we can find a sequence $\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_k$ of multi-ray clusters such that $\mathcal{C}_i$ consists of exactly r customers in $C \setminus (\mathcal{C}_1 \cup \mathcal{C}_2 \cup \cdots \mathcal{C}_{i-1})$ except the last cluster $\mathcal{C}_k$ with $r \leq |\mathcal{C}_k| \leq 2r-1$. The algorithm checks every possible sequence of the rays containing at least one customer and chooses the best one as an optimal r-gathering. Note that if a cluster is a single-ray cluster, then the algorithm skips recursive call, since it try to find an r-gathering consisting of only multi-ray clusters.

We now estimate the running time of the algorithm.

By Lemma 4.3 the depth of the recursive calls is at most $d-1$. Thus, by the tree structure of the calls, the number of calls is at most $d!$. The algorithm repeatedly constructs a multi-ray cluster with exactly r customers by Lemma 4.2. Construction of each multi-ray cluster takes $O(r \log d)$ time for each and $O(dr \log d)$ time in total. The cluster is assigned to its best facility of the cluster. The best facility of a multi-ray cluster is the nearest facility to the mid-point of p_i and the farthest customer from p_i which is on a ray except l_i. The best facility can be found in $O(d + \log m)$ time for each cluster. For each possible pair of customers we precompute the best facility. The number of such possible pairs is $(d^2r)^2$. Thus the algorithm runs in $O(d!dr \log d + (d + \log m)d^4r^2)$ time.

We can improve the running time by modifying the algorithm to store the solution of each subproblem in a table. The number of distinct subproblems is the number of the combinations of the rays checked (If we check l_1 then l_2, then the remaining subproblem is the same one which is derived if we check l_2 then l_1. Only combination matters). Thus the number of distinct subproblems is $\sum_{j=1}^{d-1} \binom{d}{i} = O(2^d)$. Then the running time is $O((d + \log m)d^4r^2 + 2^d dr \log d)$. □

Theorem 4.5. *An optimal r-gathering of C to F can be computed in $O(n + m +$*

Algorithm 3 Multi-rayClusters2(C, F, d)

Input A set C of customers, a set F of facilities on a star, and degree d of star

Output An optimal r-gathering with only multi-ray clusters if exists

if $|C| < r$ or the number of rays containing at least on customer is at most one or $F = \emptyset$ **then**
  return $\emptyset$
end if
if $|C| < 2r$ or the number of rays containing customers is two **then**
  $f \leftarrow$ the best facility for customers in C
  $A \leftarrow$ Assignment of all customers in C to f
  return A
end if
$Ans \leftarrow \emptyset$
$Best \leftarrow \infty$
for each ray l_i containing a customer **do**
  $\mathcal{C}_i \leftarrow p_i$ and his/her $r-1$ nearest customers in C /* Lemma 4.2 */
  if $\mathcal{C}_i$ is a multi-ray cluster **then**
    $A' \leftarrow$ Multi-rayClusters2($C \setminus \mathcal{C}_i, F, d$)
    if $A' \neq \emptyset$ **then**
      $f \leftarrow$ the best facility for customers in $\mathcal{C}_i$
      $A \leftarrow$ Assignment of customers in $\mathcal{C}_i$ to f and $C \setminus \mathcal{C}_i$ according to A'
      if $cost(A) < Best$ **then**
        $Best \leftarrow cost(A)$
        $Ans \leftarrow A$
      end if
    end if
  end if
end for
return Ans

$(d + \log m)d^4r^2 + d^d r^d 2^d dr \log d)$ *time.*

Proof. Similar to Theorem 3.8 we can prove the number of possible choices of the multi-ray clusters is $O(d^d r^d)$. For each choice we compute an r-gathering with Multi-rayClusters2 and compute r-gatherings of the remaining one-dimensional problems, then combine them to form an r-gathering of C to F. Then output the best one. This construction of multi-ray clusters needs $O((d+\log m)d^4r^2+2^d dr \log d)$ for each. We need not compute the best facility of each possible pair of customers for each call of Multi-rayClusters2 independently. We precompute such best facilities just once for an execution of the algorithm. Such precomputation takes $O(d^2r^2(d + \log m))$ time. We can solve all possible one dimensional r-gathering

problem in $O(n+m)$ time in total [16] and we store the solutions in a table. Note that when we solve one dimensional r-gathering problem of ray l, we may assign a cluster to the nearest facility to o located on other ray, however one can compute such f quickly. Thus the time complexity of finding an optimal r-gathering is $O(n+m+(d+\log m)d^4r^2+d^dr^d2^ddr\log d)$. ◻

If d is constant then the running time of the algorithm is polynomial.

5. Min-max-sum r-Gathering Problem

In this section we introduce a new cost function for the r-gathering problem and show that a variant of the r-gathering problem, called the min-max-sum r-gathering problem, is NP-hard even when the customers and facilities are on a star.

Let C be a set of customers, F be a set of facilities and A be an r-gathering of C to F. We define the *tree cost* of a facility f as $\sum_{c:A(c)=f} d(c,A(c))$. The *min-max-sum r-gathering problem* asks to find an r-gathering such that the maximum tree cost among all the facilities is minimum. The *decision min-max-sum r-gathering problem* is defined as follows.

Problem: Decision Min-max-sum r-Gathering Problem.

Instance: A set of customers C and a set of facilities F, an integer r, and a number q.

Question: Does there exist an r-gathering A such that for each $f \in F$, $\sum_{c:A(c)=f} d(c,A(c)) \leq q$?

We show the hardness of the decision min-max-sum r-gathering problem by reduction from the *3-partition problem* [10]. The 3-partition problem is defined as follows.

Problem: 3-Partition Problem.

Instance: A multi-set $S=\{a_1,a_2,\cdots,a_{3k}\}$ of $3k$ integers and a number b such that $\frac{b}{4}<a_i<\frac{b}{2}$ for each $1\leq i\leq 3k$ and $\sum_{a_i}=kb$.

Question: Can S be partitioned into k subsets $S_1,S_2,\cdots,S_k$ such that for each $i=1,2,\cdots,k$; $\sum_{a\in S_i} a=b$?

We now give the following theorem.

Theorem 5.1. *The decision min-max-sum r-gathering problem is NP-hard even when the customers and facilities are on a star .*

Proof. We prove the hardness of the decision min-max-sum r-gathering problem by giving a polynomial time reduction from the 3-partition problem.

Given an instance $\mathcal{I}(S,b)$ of the 3-partition problem, we construct an instance $\mathcal{I}(C,F,r,q)$ of the decision min-max-sum r-gathering problem such that $\mathcal{I}(S,b)$ has an affirmative answer if and only if $\mathcal{I}(C,F,r,q)$ has an affirmative answer. We first construct a star $\mathcal{L}=\{l_1,l_2,\cdots,l_{3k}\}$ of degree $3k$ and center o. For each $a_i\in S$ we take a customer c_i, lying on l_i, such that $d(o,c_i)=a_i$. Note that, $\frac{b}{4}<d(o,c_i)<\frac{b}{2}$ holds, since $\frac{b}{4}<a_i<\frac{b}{2}$ for each a_i. We now take $3k$ facilities $f_1,f_2,\cdots,f_{3k}$ such

that f_i is lying on ray l_i and $d(o, f_i) = \epsilon$ where $\epsilon < \min\{d(o, c_i)\}$. Finally we set $q = b + \epsilon$ and $r = 3$. In the following we prove that there is a solution to an instance $\mathcal{I}(S, b)$ if and only if $\mathcal{I}(C, F, r, q)$ has a solution.

We first assume that $\mathcal{I}(S, b)$ has an affirmative answer. Let $S_1, S_2, \cdots, S_k$ be the partition of S such that $\sum_{a \in S_i} a = b$ for each S_i. We can construct an r-gathering of instance $\mathcal{I}(C, F, r, k)$ in the following way: for each $S_i = \{a_{i_1}, a_{i_2}, a_{i_3}\}$ we assign the customers $c_{i_1}, c_{i_2}, c_{i_3}$ to the facility f_{i_1}. Note that f_{i_1} is lying on the same ray l_{i_1} as c_{i_1}. Now the cost of the facility f_{i_1} is $d(o, c_{i_1}) - \epsilon + d(o, c_{i_2}) + \epsilon + d(o, c_{i_3}) + \epsilon = b + \epsilon$. Thus each open facility f_i serves exactly 3 customers and for each $f \in F$, $\sum_{c:A(c)=f} d(c, A(c)) = b + \epsilon$.

Conversely, assume that $\mathcal{I}(C, F, r, q)$ has an affirmative answer. Let A be the corresponding r-gathering. We first claim that, each open facility in A serves exactly 3 customers. For a contradiction, assume otherwise. Let f_i be an open facility such that f_i serves at least four customers. Since $\frac{b}{4} < d(o, c_i) < \frac{b}{2}$ holds for each c_i, $\sum_{c:A(c)=f_i} d(f_i, c) \geq \sum_{c:A(c)=f_i} d(o, c) + 2\epsilon > 4 \times \frac{b}{4} + 2\epsilon = b + 2\epsilon$, a contradiction. We now claim that, $\sum_{c:A(c)=f_i} d(f_i, c) = b + \epsilon$ holds for each open facility f_i. Assume for a contradiction that, $\sum_{c:A(c)=f_i} d(f_i, c) < b + \epsilon$ holds for an open facility f_i. Thus we get $\sum_{c:A(c)=f_i} d(o, c) < b$. Let C' be the set of customers assigned to some facility other than f_i. Clearly $|C'| = 3k - 3$. Since $\sum_{c \in C} d(o, c) = kb$ and $\sum_{c:A(c)=f_i} d(o, c) < b$, we get $\sum_{c \in C'} d(o, c) > (k-1)b$. Then there is at least one facility f_j for which $\sum_{c:A(c)=f_j} d(o, c) > b$ holds. Thus $\sum_{c:A(c)=f_j} d(f_j, c) > b + \epsilon$. A contradiction. ◻

6. One-dimensional Uncertain r-Gathering Problem

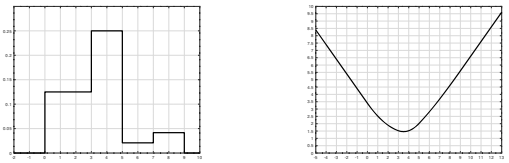
In this section we give two algorithms for the uncertain r-gathering problem on a line.

Let $C = \{c_1, c_2, \cdots, c_n\}$ be a set of n uncertain customer on a horizontal line where the location of each customer c_i is specified by his/her PDF $g_i : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the horizontal line. We consider the facilities are ordered from left to right. We sometime regard f_i as its coordinate.

6.1. *Histogram*

In this section we give an algorithm for the uncertain r-gathering problem when the location of each customer is specified by a piecewise uniform function, i.e., a histogram.

We consider the PDF of each customer c_i is defined as a piecewise uniform function g_i, i.e., a histogram. The PDF of each uncertain customer is independent. We consider the histogram model since it can be used to approximate any PDF [1]. The histogram model is considered by Wang and Zhang [20] for the uncertain k-center problem on a line. Each g_i consists of at most $k+1$ pieces where each piece is a uniform function. Each customer c_i has $k+2$ points $x_{i0}, x_{i1}, \cdots, x_{i(k+1)}$ on the



(a) (b)

Fig. 4. (a) Illustration of a histogram with 6 pieces and (b) corresponding function of expected distance.

line, where $x_{i0} < x_{i1} < \cdots < x_{i(k+1)}$, and $k+1$ values $y_{i0}, y_{i1}, \cdots, y_{ik}$ such that $g_i(x) = y_{ij}$ if $x_{ij} \leq x < x_{i(j+1)}$. We consider $x_{i0} = -\infty$, $x_{i(k+1)} = \infty$, $y_{i0} = 0$, and $y_{ik} = 0$. Figure 4(a) illustrates a histogram of 6 pieces. The expected distance $E[d(p, c_i)]$ from a point p to c_i is defined as follows.

$$E[d(p, c_i)] = \int_{-\infty}^{\infty} g_i(x)|x - p|dx$$

A function $h : \mathbb{R} \to \mathbb{R}$ is called a *unimodal function* if there is a point p such that $h(x)$ is monotonically decreasing in $(-\infty, p]$ and monotonically increasing in $[p, \infty)$. Wang and Zhang gave the following lemma [20].

Lemma 6.1 ([20]). *Let c_i be an uncertain customer on a line which is specified by a histogram of $k+1$ pieces. Then the function $E[d(p, c_i)]$ for $p \in \mathbb{R}$ is a unimodal function consisting of a parabola in each interval $[x_{ij}, x_{i(j+1)})$. Furthermore the function $E[d(p, c_i)]$ can be explicitly computed in $O(k)$ time.*

Outline of Proof Without loss of generality, assume that $x_{it} \leq p \leq x_{i(t+1)}$. Then the function $E[d(p, c_i)]$ can be written as follows [20].

$$\begin{aligned} E[d(p, c_i)] &= y_{it}p^2 \\ &+ \left[\sum_{j=0}^{t-1} y_{ij}\left(x_{i(j+1)} - x_{ij}\right) - \sum_{j=t+1}^{k} y_{ij}\left(x_{i(j+1)} - x_{ij}\right) - y_{it}(x_{it} + x_{i(t+1)})\right] p \\ &+ \frac{1}{2}\left[\sum_{j=t+1}^{k} y_{ij}\left(x_{i(j+1)}^2 - x_{ij}^2\right) - \sum_{j=0}^{t-1} y_{ij}\left(x_{i(j+1)}^2 - x_{ij}^2\right) + y_{it}(x_{it}^2 + x_{i(t+1)}^2)\right] \end{aligned} \tag{6.1}$$

Thus we can write $E[d(p, c_i)]$ as $a_{i1}(t)p^2 + a_{i2}(t)p + a_{i3}$ where each of $a_{i1}(t), a_{i2}(t)$, and $a_{i3}(t)$ depends on t satisfying $x_{it} \leq p \leq x_{i(t+1)}$. Note that if $y_{it} = 0$ then the function $E[d(p, c_i)]$ is a straight line in the interval $[x_{it}, x_{i(t+1)})$ which we consider as a special parabola. Figure 4(b) illustrates the $E[d(p, c_i)]$ function for the histogram in Figure 4(a). We can compute the co-efficients $a_{i1}(j)$ for all j in $O(k)$ time. Moreover, the summation terms in $a_{i2}(j)$ and $a_{i3}(j)$ for all j can be computed in

$O(k)$ time in total. Thus for all j, we can compute the $a_{i2}(j)$ and $a_{i3}(j)$ in $O(k)$ time. Hence the function $E[d(p, c_i)]$ can be computed explicitly in $O(k)$ time. □

We now give the following lemma.

Lemma 6.2. *Let c_i be an uncertain customer on a line which is specified by a histogram of $k+1$ pieces, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the line. We can compute the expected distances between all facilities and the uncertain customer in $O(m+k)$ time. Furthermore the expected distances between the facilities and the uncertain customer can be sorted in $O(m)$ time.*

Proof. We first precompute the co-efficients $a_{i1}(j), a_{i2}(j), a_{i3}(j)$ of function $E[d(p, c_i)]$ for all j in $O(k)$ time by Lemma 6.1. With the precomputed function $E[d(p, c_i)]$, the expected distance between the uncertain customer and a facility f_u can be computed in $O(\log k)$ time using binary search to find the $[x_{it}, x_{i(t+1)})$ where f_u is located. Thus the expected distance between all facilities and the uncertain customer can be computed in $O(m \log k)$ time. However, we can improve the running time to $O(m + k)$ by a plane sweep from left to right. We take the facilities from left to right, determine the corresponding interval $[x_{ij}, x_{i(j+1)})$, and compute the expected distance. Since both the facilities and the $x_{i1}, x_{i2}, \cdots, x_{ik}$ are ordered from left to right, the search for the interval in which f_u is located can start from the interval in which f_{u-1} is located. Hence each x_{ij} will be considered once. Thus the total running time is $O(m + k)$. We now show that the sorted list of the expected distances between the facilities and the uncertain customer can be constructed in $O(m + k)$ time. Since $E[d(p, c_i)]$ is a unimodal function, there is a facility f_u such that $E[d(f_{v-1}, c_i)] \geq E[d(f_v, c_i)]$ for any $1 < v \leq u$, and $E[d(f_v, c_i)] \leq E[d(f_{v+1}, c_i)]$ for any $u \leq v < m$. Thus we have a descending list of expected distances for $f_1, f_2, \cdots, f_u$ and ascending list of expected distances for $f_{u+1}, f_{u+2}, \cdots, f_m$. We can merge these two lists into an ascending list of expected distances in $O(m)$ time. □

Corollary 6.3. *Let $C = \{c_1, c_2, \cdots, c_n\}$ be set of n uncertain customers on a line each of which is specified by a histogram of $k+1$ pieces, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the line. The expected distances between all pair of uncertain customers and facilities can be computed and sorted in $O(nk + mn \log n)$ time.*

Proof. By Lemma 6.2, we can compute n sorted list of expected distances between customers and facilities in $O(nk + mn)$ time. The n sorted lists can be merged into a single list using min-heap in $O(mn \log n)$ time. □

We first consider the decision version of the uncertain r-gathering problem on a line. Given a set of uncertain customers C, a set of facilities F on a line, and a number b, the decision uncertain r-gathering problem asks to determine whether there is an r-gathering A of C to F such that $E[d(c, A(c))] \leq b$ for each $c \in C$. The following lemma is known [20].

Lemma 6.4 ([20]). *Let c be an uncertain customer on a line which is specified by a histogram of $k+1$ pieces, and b be a number. Then the points p for which $E[d(c,p)] \leq b$ holds form an interval on the line.*

We call the interval which admits $E[d(c,p)] \leq b$ for customer c a (c,b)-interval and denote the interval by $[s_b(c), t_b(c)]$. Furthermore, in any r-gathering A with cost at most b, $A(c)$ is in $[s_b(c), t_b(c)]$. Thus to find whether there is an r-gathering satisfying $E[d(c,p)] \leq b$ for each customer c, it is sufficient to solve the following problem. Given a set of facilities F on a line and a set of customers C where each customer $c \in C$ has an interval $[s(c), t(c)]$ on the line, the *interval r-gathering problem* asks to determine whether there is an r-gathering A such that each facility $f \in F$ serves zero or at least r customers and for each customer $c \in C$, $s(c) \leq A(c) \leq t(c)$ holds.

We now give an algorithm for the interval r-gathering problem. Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities and $C = \{c_1, c_2, \cdots, c_n\}$ be a set of customers on a line where each customer c_i has an interval $I_i = [s(c_i), t(c_i)]$. An interval I_i is called the *leftmost interval* if for each $c_j \neq c_i$, $t(c_i) \leq t(c_j)$ holds, and the customer c_i is called the *leftmost customer*. We call a customer c_i is in left with respect to customer c_j if and only if $t(c_i) \leq t(c_j)$. A facility f_u is called the *preceding facility* of c_i if $s(c_i) \leq f_u \leq t(c_i)$ and there is no facility f_v such that $f_u < f_v \leq t(c_i)$. Similarly a facility f_u is called the *following facility* of c_i if $s(c_i) \leq f_u \leq t(c_i)$ and there is no facility f_v such that $s(c_i) \leq f_v < f_u$. We call a customer c_j a *right neighbor* of c_i if $t(c_j) \geq t(c_i)$ and $s(c_j) \leq t(c_i)$.

Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities and $C = \{c_1, c_2, \cdots, c_n\}$ be a set of customers on a line where each customer c_i has an interval I_i. Let c_i be the leftmost customer, f_u be the preceding facility of c_i, and P_u be the set of customers containing f_u in their intervals, i.e., potential customers that can be assigned to f_u. We have the following lemmas.

Lemma 6.5. *If there is an interval r-gathering of C to F, then there is an interval r-gathering where f_u is the leftmost open facility. Furthermore, the customers assigned to f_u have consecutive right end-points in P_u including c_i.*

Proof. We first prove that there is an interval r-gathering where f_u is the leftmost open facility. Assume for a contradiction that there is no interval r-gathering where f_u is the leftmost open facility. Let A be an interval r-gathering where $f_v \neq f_u$ is the leftmost open facility. We can observe that $f_v \leq f_u$, since in each interval r-gathering, c_i is assigned to a facility within the interval I_i and f_u is the preceding facility of c_i. Let $\mathcal{C}_v$ be the set of customers assigned to f_v in A. For any customer c_j in $\mathcal{C}_v$, we have $s(c_j) \leq f_v \leq f_u \leq t(c_i) \leq t(c_j)$, since I_i is the leftmost interval. We now derive a new interval r-gathering by reassigning the customers $\mathcal{C}_v$ to f_u, a contradiction.

We now prove that the customers assigned to f_u have consecutive right end-points in P_u. We call a pair $c_j, c_k \in P_u$ a reverse pair if $t(c_j) < t(c_k)$, c_k assigned to

f_u, and c_j assigned to $f_v > f_u$. Assume for a contradiction that there is no interval r-gathering where the customers assigned to f_u have consecutive right end-points in P_u. Let A' be an interval r-gathering with minimum number of reverse pairs but the number is not zero. Let c_j, c_k be a reverse pair in A' where $t(c_j) < t(c_k)$, and c_j is assigned to facility f_w, and c_k is assigned to f_u. Since $t(c_k) > t(c_j)$ and $f_w \geq f_u$, we get $s(c_k) \leq f_w \leq t(c_k)$. We now derive a new interval r-gathering with less reverse pairs by reassigning c_j to f_u and c_k to f_w, a contradiction. □

Lemma 6.6. *Let c_j be the leftmost customer in $C \setminus P_u$, and $P'_u \subseteq P_u$ be the customers such that for each $c \in P'_u$, $t(c) < t(c_j)$. If there is an interval r-gathering, then there is an interval r-gathering satisfying one of the following.*
(a) If $|P'_u| < r$, then the customers assigned to f_u are the r leftmost customers in P_u.
(b) If $|P'_u| \geq r$, then $\max\{|P'_u| - r + 1, r\}$ leftmost customers of P_u are assigned to f_u (possibly with more customers).

Proof. (a) By Lemma 6.5, there is an interval r-gatering where the customers assigned to f_u have consecutive right end-points in P_u and the leftmost customer c_i is assigned to f_u. Thus the leftmost r customers P^l_u in P_u are assigned to f_u. We now prove that there is an interval r-gathering where no customer in $P_u \setminus P^l_u$ is assigned to f_u. Assume for a contradiction that in every interval r-gathering there are some customers in $P_u \setminus P^l_u$ which are assigned to f_u. Let A be an interval r-gathering where the number of customers in $P_u \setminus P^l_u$ assigned to f_u is minimum, and c_k be a customer in $P_u \setminus P^l_u$ which is assigned to f_u. Since $|P'_u| < r$, we get $t(c_k) > t(c_j)$. Let c_j is assigned to f_v in A. We now derive a new r-gathering by reassigning c_k to f_v, a contradiction.
(b) We first consider $r \leq |P'_u| < 2r$. In this case $\max\{|P'_u| - r + 1, r\} = r$. Hence by Lemma 6.5 the leftmost r customers in P_u are assigned to f_u.
We now consider $|P'_u| \geq 2r$. In this case, $\max\{|P'_u| - r + 1, r\} = |P'_u| - r + 1$. Let P''_u be the leftmost $|P'_u| - r + 1$ customers in P_u. Note that, by the definition of P'_u, P''_u are also the $|P'_u| - r + 1$ leftmost customers in P'_u. Assume for a contradiction that there is no interval r-gathering where P''_u are assigned to f_u. Let A' be an interval r-gathering with the maximum number of customers $Q_u \subset P''_u$ assigned to f_u. Let $c_k \in P''_u$ be the customer with smallest $t(c_k)$ which is not assigned to f_u. Let c_k is assigned to $f_v \geq f_u$. By Lemma 6.5, any customer $c_x \in P''_u$ with $t(c_x) \geq t(c_k)$ is not assigned to f_u. Let $\mathcal{C}_v$ be the customers assigned to f_v. We first claim that $|\mathcal{C}_v| = r$, otherwise we can reassign c_k to f_u, contradicting our assumption. We now claim that $\mathcal{C}_v$ consists of r customers with consecutive right end-points in P_u. Assume otherwise for a contradiction. Let A'' be an interval r-gathering with the minimum number of reverse pairs where a reverse pair is a pair of customer c_x, c_y with $t(c_x) \leq t(c_y)$, c_y assigned to f_v, c_x assigned to $f_w > f_v$. Since $t(c_x) \leq t(c_y)$ and $f_v \leq f_w$, we get $s(c_y) \leq f_w \leq t(c_y)$. We now derive a new interval r-gathering by reassigning c_x to f_v and c_y to f_w, a contradiction. Now since $|Q_u| < |P'_u| - r + 1$,

we get $|P'_u \setminus Q_u| \geq r$. Thus $\mathcal{C}_v \subset P'_u \subseteq P_u$. We now derive a new interval r-gathering by assigning $\mathcal{C}_v$ to f_u. A contradiction. □

We now give an algorithm **Interval-*r*-gather** for the interval r-gathering problem.

We have the following theorem.

Theorem 6.7. *The algorithm **Interval-r-gather** decides whether there is an interval r-gathering of C to F, and constructs one if it exists in $O(m+n\log n+nr^{\frac{n}{r}})$ time.*

Proof. The correctness of Algorithm Interval-r-gather is immediate from lemma 6.5 and 6.6. We check every possible cases with backtracking.

We now estimate the running time of the algorithm. We can sort the customers based on their right end-points in $O(n\log n)$ time. For each customer we can precompute the preceding facility f_u in $O(n+m)$ time in total. For each facility f_u we can precompute the sets of customers P_u containing each facility and the leftmost customer c_j having left end-point on right of f_u in $O(n+m)$ time in total. In each call to Interval-r-gather, we need $O(|P_u|)$ time and at most r recursive calls to Interval-r-gather. Let $T(n)$ be the running time of the algorithm for n customers. We have $T(n) \leq O(|P_u|) + \sum_{i=1}^{r} T(n-r+1) \leq O(nr^{\frac{n}{r}})$. Thus the running time of the algorithm is $O(m+n\log n+nr^{\frac{n}{r}})$. □

We now have the following theorem.

Theorem 6.8. *Let $C = \{c_1, c_2, \cdots, c_n\}$ be a set of uncertain customers on a line each of which is specified by a histogram consisting of $k+1$ pieces, and $F = \{f_1, f_2, \cdots, f_m\}$ be a set of m facilities on the line. Then the optimal r-gathering can be constructed in $O(nk + mn\log n + (m + n\log kn + nr^{\frac{n}{r}})\log mn)$ time.*

Proof. We give outline of an algorithm to compute optimal r-gathering. We first compute the $E[d(p, c_i)]$ function for each $c_i \in C$. This takes $O(nk)$ time in total by Lemma 6.1. By Corollary 6.3, we compute the sorted list of all expected distances between customers and facilities in $O(nk + mn\log n)$ time. We find the optimal r-gathering by binary search, using the $O(m+n\log n+nr^{\frac{n}{r}})$ time algorithm for the interval r-gathering problem $\log mn$ times. For each r-interval gathering problem, we compute the (c_i, b)-intervals in $O(n\log k)$ time. Thus finding optimal r-gathering by binary search requires $O(nk + mn\log n + (m + n\log kn + nr^{\frac{n}{r}})\log mn)$ time. □

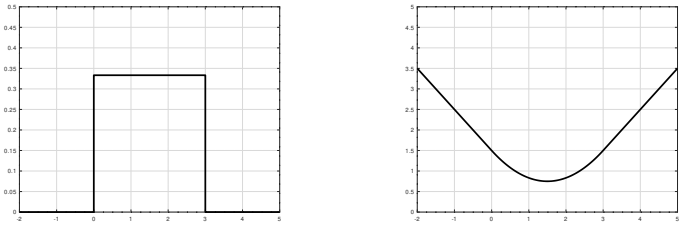
6.2. *Uniform Distribution*

In this section we give an algorithm for the uncertain r-gathering problem when each customer location is specified by a well-separated uniform distribution.

Algorithm 4 Interval-r-gather(C, F)

if $|C| < r$ or $F = \emptyset$ **then**
  return $\emptyset$
end if
$c_i \leftarrow$ leftmost customer in C
$f_u \leftarrow$ preceding facility of c_i
$P_u \leftarrow$ the set of customers containing f_u in their intervals
$c_j \leftarrow$ leftmost customer in $C \setminus P_u$
$P'_u \leftarrow$ the set of customers in P_u having smaller right end-point than $t(c_j)$
$F' \leftarrow$ the set of facilities right to f
if $|P_u| < r$ **then**
  return $\emptyset$
end if
if $|P'_u| < r$ **then**
  $\mathcal{C}_u \leftarrow$ the set of r leftmost customers in P_u /* Lemma 5(a) */
  $A \leftarrow$ Assignment of $\mathcal{C}_u$ to f_u
  $Ans \leftarrow$ Interval-r-gather$(C \setminus \mathcal{C}_u, F')$
  if $Ans \neq \emptyset$ **then**
    return Assignment of $\mathcal{C}_u$ and $C \setminus \mathcal{C}_u$ by A and Ans, respectively
  end if
  return $\emptyset$
end if
$\mathcal{C}_u \leftarrow$ the set of $\max\{r, |P'_u| - r + 1\}$ leftmost customers in P_u /* Lemma 5(b) */
$A \leftarrow$ Assignment of $\mathcal{C}_u$ to f_u
$P''_u \leftarrow P'_u \setminus \mathcal{C}_u$
while P''_u is not empty **do**
  $Ans \leftarrow$ Interval-r-gather$(C \setminus \mathcal{C}_u, F')$
  if $Ans \neq \emptyset$ **then**
    return Assignment of $\mathcal{C}_u$ and $C \setminus \mathcal{C}_u$ by A and Ans, respectively
  end if
  $c_k \leftarrow$ leftmost customer in P''_u /* (possibly with more customers) */
  $\mathcal{C}_u \leftarrow \mathcal{C}_u \cup \{c_k\}$
  $A \leftarrow$ Assignment of $\mathcal{C}_u$ to f_u
  $P''_u \leftarrow P''_u \setminus \{c_k\}$
end while
return $\emptyset$

In the uniform distribution model, location of each customer c_i is specified by a probability density function $g_i : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$ where $g_i(p) = 1/(t_i - s_i)$ if $s_i \leq p \leq t_i$ and $g_i(p) = 0$ otherwise. We denote the uniform distribution between $[s_i, t_i]$ by $U(s_i, t_i)$. The customer c_i having a uniform distribution $U(s_i, t_i)$ is denoted



(a) (b)

Fig. 5. (a) Illustration of a uniform distribution and (b) corresponding function of expected distance.

by $c_i \sim U(s_i, t_i)$. Figure 5(a) illustrates a uniform distribution where $s_i = 0$ and $t_i = 3$. The range of $U(s_i, t_i)$, denoted by e_i, is the value of $t_i - s_i$, and the mean of $U(s_i, t_i)$, denoted by μ_i, is the value of $\frac{s_i+t_i}{2}$. The uniform distribution model is a special case of the histogram model described in Section 6.1. We now have the following lemma.

Lemma 6.9. *Let $c \sim U(s, t)$ be an uncertain customer. Then the function $E[d(p, c)]$ consists of a parabola in the interval $[s, t]$ and two straight lines of slope +1 and -1 in interval (t, ∞) and $(-\infty, s)$, respectively. Furthermore, the minimum value of $E[d(p, c)]$ is $\frac{e}{4}$ and the value of $E[d(p, c)]$ at s, t is $\frac{e}{2}$.*

Proof. We use the Equation 6.1 to compute the function $E[d(p, c)]$.

$$E[d(p,c)] = \begin{cases} \mu - p & \text{if } p < s \\ \frac{1}{e}(p-\mu)^2 + \frac{e}{4} & \text{if } s \le p \le t \\ -\mu + p & \text{if } p > t \end{cases} \tag{6.2}$$

At $p = s$ we get $E[d(s, C)] = \frac{1}{t-s}\left(s - \frac{s+t}{2}\right)^2 + \frac{t-s}{4} = \frac{t-s}{2} = \frac{e}{2}$. Similarly, $E[d(t, C)] = \frac{e}{2}$. Now for $p < s$ and $p > t$, $E[d(p, c)] \ge \frac{t-s}{2}$. The minimum value of the parabola $\frac{1}{t-s}\left(p - \frac{s+t}{2}\right)^2 + \frac{t-s}{4}$ is $\frac{e}{4}$ at $p = \frac{s+t}{2}$. □

We have the following lemma.

Lemma 6.10. *Let $c \sim U(s, t)$ be an uncertain customer and b be a number. Then the (c, b)-interval can be computed in $O(1)$ time.*

Proof. To find the (c, b)-interval, we first compute the inverse of the Equation 6.2. For $E[d(p, c)] = b > \frac{e}{2}$, we have $p < s$ or $p > t$. Thus we get, $p = \mu \pm b$. For $\frac{e}{4} \le E[d(p, c)] = b \le \frac{e}{2}$, we have $s \le p \le t$. Thus we get $p = \mu \pm \sqrt{l(b - \frac{e}{4})}$. Finally there is no p for which $E[d(p, C)] < \frac{e}{4}$. Hence the (c, b)-interval for $b < \frac{e}{4}$ is empty.

Thus the (c, b)-interval I can be written as following.

$$I = \begin{cases} [\mu - b, \mu + b] & \text{if } b > \frac{e}{2} \\ [\mu - \sqrt{l(b - \frac{e}{4})}, \mu + \sqrt{l(b - \frac{e}{4})}] & \text{if } \frac{e}{4} \leq b \leq \frac{e}{2} \\ \emptyset & \text{if } b < \frac{e}{4} \end{cases} \tag{6.3}$$

By Equation 6.3 we can compute (c, b)-interval in $O(1)$ time. □

Let $c_i \sim U(s_i, t_i), c_j \sim U(s_j, t_j)$ be two uncertain customers. Let $e_{max} = \max\{e_i, e_j\}$ and $e_{min} = \min\{e_i, e_j\}$. We call c_i, c_j *well-separated* if none of the intervals $[s_i, t_i]$ and $[s_j, t_j]$ is contained within the other and $|\mu_i - \mu_j| \geq \frac{1}{2}\sqrt{e_{min}(e_{max} - e_{min})}$.

Lemma 6.11. *Let $c_i \sim U(s_i, t_i), c_j \sim U(s_j, t_j)$ be two uncertain well-separated points and b be a number. Let I_i, I_j be the (c_i, b)-interval and (c_j, b)-interval respectively. Then none of I_i and I_j is properly contained in the other.*

Proof. Since c_i and c_j are well-separated, it is easy to observe that $s_i = s_j$ if and only if $t_i = t_j$. In this case the claim trivially holds. We thus consider otherwise. Without loss of generality we assume that, $s_i < s_j$, $t_i < t_j$ and $e_i \leq e_j$. Since $s_i < s_j$ and $t_i < t_j$, we get $\mu_i = \frac{s_i + t_i}{2} < \frac{s_j + t_j}{2} = \mu_j$. We now have two cases.

Case 1: $e_j \geq 2e_i$. In this case we have three subcases to consider.

Case 1a: $b > \frac{e_j}{2}$.

By Equation 6.3 we get $I_i = [\mu_i - b, \mu_i + b]$. Similarly, we get $I_j = [\mu_j - b, \mu_j + b]$. Now since $\mu_i < \mu_j$, we get $\mu_i - b < \mu_j - b$ and $\mu_i + b < \mu_j + b$. Thus none of I_i, I_j is contained within the other.

Case 1b: $\frac{e_j}{4} \leq b \leq \frac{e_j}{2}$.

Since $e_j \geq 2e_i$, we have $\frac{e_i}{2} \leq \frac{e_j}{4}$. By Equation 6.3, $I_i = [\mu_i - b, \mu_i + b]$ and $I_j = [\mu_j - \sqrt{e_j(b - e_j/4)}, \mu_j + \sqrt{e_j(b + e_j/4)}]$. Assume for a contradiction that either I_i or I_j is contained within the other. We first consider I_i is contained within I_j. Since $b > \frac{e_i}{2}$, we get $\mu_i - b < s_i$ and $\mu_i + b > t_i$. On the other hand, since $b \leq \frac{e_j}{2}$ we get $\mu_j - \sqrt{e_j(b - e_j/4)} \geq s_j$ and $\mu_j - \sqrt{e_j(b - e_j/4)} \leq t_j$. Now since I_i is contained within I_j, we have $\mu_j - \sqrt{e_j(b - e_j/4)} \leq \mu_i - b$ and $\mu_j + \sqrt{e_j(b - e_j/4)} \geq \mu_i + b$. Thus we get $s_j < s_i$ and $t_i < t_j$, a contradiction.

We now consider I_j is contained within I_i. In this case, $\mu_i - b \leq \mu_j - \sqrt{e_j(b - e_j/4)}$ and $\mu_i + b \geq \mu_j + \sqrt{e_j(b - e_j/4)}$. Note that, the absolute value of the slope of tangent of parabola $\frac{1}{e_j}(p - \mu_j)^2 + \frac{e_j}{4}$ at any point $p \in [s_j, t_j]$ is less than 1. Hence the interval I_j at $b = \frac{e_j}{4}$ must be contained within the interval I_i at $b = \frac{e_j}{4}$. At $b = \frac{e_j}{4}$, we have $I_i = [\mu_i - \frac{e_j}{4}, \mu_i + \frac{e_j}{4}]$ and $I_j = [\mu_j, \mu_j]$. Since I_j is contained within I_i, we get $\mu_j - \mu_i \leq \frac{e_j}{4}$. Now since I_i is not contained within I_j and $e_j \geq 2e_i$, we get $s_j = \mu_j - \frac{e_j}{2} \leq \mu_i + \frac{e_j}{4} - \frac{e_i}{2} = \mu_i - \frac{e_i}{4} \leq \mu_i - \frac{e_i}{2} = s_i$

Similarly, we can show $t_j \geq t_i$. Thus I_i is contained within I_j, a contradiction.

Case 1c: $b < \frac{e_j}{4}$.

Since $b < \frac{e_j}{4}$, we get $I_j = \emptyset$. Thus the claim holds.

Case 2: $e_j < 2e_i$. For this case, we have four subcases to consider.
Case 2a: $b > \frac{e_j}{2}$. Similar to Case 1a.
Case 2b: $\frac{e_i}{2} < b \leq \frac{e_j}{2}$.
In this case by Equation 6.3, $I_i = [\mu_i - b, \mu_i + b]$ and $I_j = [\mu_j - \sqrt{e_j(b - e_j/4)}, \mu_j + \sqrt{e_j(b + e_j/4)}]$. Assume for a contradiction that either I_i or I_j is contained within the other. We first consider I_i is contained within I_j. Since $b > \frac{e_i}{2}$, we get $\mu_i - b < s_i$ and $\mu_i + b > t_i$. On the other hand, since $b \leq \frac{e_j}{2}$ we get $\mu_j - \sqrt{e_j(b - e_j/4)} \geq s_j$ and $\mu_j - \sqrt{e_j(b - e_j/4)} \leq t_j$. Now since I_i is contained within I_j, we have $\mu_j - \sqrt{e_j(b - e_j/4)} \leq \mu_i - b$ and $\mu_j + \sqrt{e_j(b - e_j/4)} \geq \mu_i + b$. Thus we get $s_j < s_i$ and $t_i < t_j$, a contradiction.
We now consider I_j is contained within I_i. In this case, $\mu_i - b \leq \mu_j - \sqrt{e_j(b - e_j/4)}$ and $\mu_i + b \geq \mu_j + \sqrt{e_j(b - e_j/4)}$. Since the absolute value of the slope of tangent of parabola $\frac{1}{e_j}(p - \mu_j)^2 + \frac{e_j}{4}$ at any point $p \in [s_j, t_j]$ is less than 1, the interval I_j at $b = \frac{e_i}{2}$ must be contained within the interval I_i at $b = \frac{e_i}{2}$. At $b = \frac{e_i}{4}$, we have $I_i = [\mu_i - \frac{e_i}{2}, \mu_i + \frac{e_i}{2}]$ and $I_j = [\mu_j - \sqrt{e_j(\frac{e_i}{2} - \frac{e_j}{4})}, \mu_j + \sqrt{e_j(\frac{e_i}{2} - \frac{e_j}{4})}]$. Thus I_j is contained within I_i if and only if $\mu_j - \mu_i \leq \frac{e_i}{2} - \frac{1}{2}\sqrt{e_j(\frac{e_i}{2} - \frac{e_j}{4})}$. Now since $e_i \leq e_j$, we have $2e_i - e_j \leq e_j$. Hence we get,

$$\frac{e_i}{2} - \frac{1}{2}\sqrt{e_j(\frac{e_i}{2} - \frac{e_j}{4})} \leq \frac{e_i}{2} - -\frac{1}{2}\sqrt{(\frac{e_i}{2} - \frac{e_j}{4})^2} = \frac{e_j - e_i}{2}$$

Thus I_j is contained within I_i if and only if $\mu_j - \mu_i \leq \frac{e_j - e_i}{2}$. Now since, none of I_i, I_j is contained within the other, we have $\mu_j - \mu_i > \frac{e_j - e_i}{2}$.
Case 2c: $\frac{e_j}{4} \leq b \leq \frac{e_i}{2}$. In this case, $I_i = [\mu_i - \sqrt{e_i(b - e_i/4)}, \mu_i + \sqrt{e_i(b + e_i/4)}]$ and $I_j = [\mu_j - \sqrt{e_j(b - e_j/4)}, \mu_j + \sqrt{e_j(b + e_j/4)}]$. Assume for a contradiction that I_i or I_j is contained within the other. We first consider I_i is contained within I_j. Since $\mu_i \leq \mu_j$, I_i is contained within I_j if and only if $\mu_i - \sqrt{e_i(b - e_i/4)} \geq \mu_j - \sqrt{e_j(b - e_j/4)}$ which yields $\mu_j - \mu_i \leq \sqrt{e_j(b - e_j/4)} - \sqrt{e_i(b - e_i/4)}$. Similarly, I_j is contained within I_i if and only if $\mu_j - \mu_i \leq \sqrt{e_i(b - e_i/4)} - \sqrt{e_j(b - e_j/4)}$. Thus either I_i or I_j is contained within the other if and only if $\mu_j - \mu_i \leq |\sqrt{e_j(b - e_j/4)} - \sqrt{e_i(b - e_i/4)}|$.
Let $h(b) = \sqrt{e_j(b - e_j/4)} - \sqrt{e_i(b - e_i/4)}$. We now show that, the function $h(b)$ is increasing at any point $b \geq e_j/4$. Clearly, $h(b)$ is not defined for $b < e_j/4$. We can calculate the derivative of $h(b)$ as follows.

$$\frac{d}{db}h(b) = \sqrt{\frac{e_j}{4b - e_j}} - \sqrt{\frac{e_i}{4b - e_i}} = \frac{\sqrt{e_j(4b - e_i)} - \sqrt{e_i(4b - e_j)}}{\sqrt{(4b - e_j)(4b - e_i)}}.$$

Since $e_i \leq e_j$, we get $\sqrt{e_j(4b - e_i)} \geq \sqrt{e_i(4b - e_j)}$. Thus $\frac{d}{db}h(b) > 0$ for any $b \geq e_j/4$, and hence the function $h(b)$ is increasing. We now show that the maximum value of $|h(b)|$ within interval $[]\frac{e_j}{4}, \frac{e_i}{2}]$ is at $b = e_j/4$. We first observe that $h(b) = 0$ at $b = \frac{e_i + e_j}{4}$. Since $e_j \geq e_i$, $\frac{e_j + e_i}{4} \geq \frac{e_i}{2}$. Thus $|h(b)|$ is decreasing in the interval $[\frac{e_j}{4}, \frac{e_i}{2}]$. Hence the maximum value of $|h(b)|$ within the interval $[\frac{e_j}{4}, \frac{e_i}{2}]$ is at $b = \frac{e_j}{4}$. Thus the maximum value of $|h(b)|$ is

$$\left|h\left(\frac{e_j}{4}\right)\right| = \left|\sqrt{e_j\left(\frac{e_j}{4}-\frac{e_j}{4}\right)} - \sqrt{e_i\left(\frac{e_j}{4}-\frac{e_i}{4}\right)}\right| = \frac{1}{2}\sqrt{e_i\left(e_j-e_i\right)}$$

Since c_i, c_j are well-separated, $\mu_j - \mu_i$ cannot be greater than $\frac{1}{2}\sqrt{e_i(e_j-e_i)}$, a contradiction.

Case 2d: $b < \frac{e_j}{4}$. Similar to Case 1c. □

If the customer locations are specified by well-separated uniform distributions, we can solve the decision version of uncertain r-gathering problem by dynamic programming as follows. A subproblem asks to determine whether there is an r-gathering with cost at most b for the set of customers $c_1, c_2, \cdots, c_i$. Thus we have at most n distinct subproblems, and to solve a subproblem we need to check n smaller subproblems, so we can design an $O(m + n^2)$ time algorithm.

We can improve the running time as follows. A subproblem $SP(i)$ asks to find a set of customers C_i and an interval r-gathering A of customers $C_i \subseteq C$ to $F_i = \{f_1, f_2, \cdots, f_i\}$ such that (1) C_i contains every customer c_i with $t(c_i) \leq f_i$ (possibly with more customers), (2) f_i serves at least r customers, and (3) $\max_{c\in C_i}\{t(c)\}$ is minimum. Let $c_{z(i)}$ be the customer with $\max_{c\in C_i}\{t(c)\}$. We can observe that there is a proper interval r-gathering of C to F if and only if some $SP(i)$ with $f_i \geq s(c_n)$ has a solution.

Lemma 6.12. *If $SP(i)$ has a solution, then there is an interval r-gathering where customers assigned to each open facility have consecutive right end-points.*

Proof. In an interval r-gathering A we call a pair of customers c_u, c_v a reverse pair, if $t(c_u) < t(c_v)$ and $A(c_u) \geq A(c_v)$. Let C_i be the set of customers corresponding to $SP(i)$. For a contradiction, assume that there is no interval r-gathering where customers assigned to each open facility is consecutive. Let A_i be an interval r-gathering corresponding to $SP(i)$ with minimum number of reverse pairs. Let c_u, c_v be a reverse pair. Since all the intervals are proper, $s(c_u) < s(c_v)$. Thus we have $s(c_u) \leq A_i(c_v) \leq t(c_u)$, and $s(c_v) \leq A_i(c_u) \leq t(c_v)$. Now we can derive a new r-gathering by reassigning c_u to $A_i(c_v)$ and c_v to $A_i(c_u)$, which reduces the number of reverse pairs by one. A contradiction. □

We now have the following lemma.

Lemma 6.13. *If $SP(i)$ and $SP(j)$ have solutions and $i < j$, then $t(c_{z(i)}) \leq t(c_{z(j)})$.*

Proof. For a contradiction assume that $t(c_{z(i)}) > t(c_{z(j)})$. Let A_j be an interval r-gathering corresponding to $SP(j)$. Since all the intervals are proper, we have $s(c_{z(i)}) > s(c_{z(j)})$, and $s(c_{z(j)}) \leq f_i$. Let $\mathcal{C}_j$ be the set of customers assigned to any facility between f_i to f_j (including f_i, f_j) in A_j. For any customer $c_k \in \mathcal{C}_j$, we have $s(c_k) \leq f_i$ and $t(c_k) \geq f_i$. We now derive a new interval r-gathering A'_j by

reassigning the leftmost r customers $\mathcal{C}_j$ to f_i. Clearly, $\max_{c\in\mathcal{C}_j}\{t(c)\} < t(c_{z(i)})$ and thus A'_j is a solution of $SP(i)$, a contradiction. □

Using Lemma 6.12 and 6.13, we can determine whether $SP(i)$ has a solution or not. We have two cases. If $f_i \leq t(c_1)$, then $SP(i)$ may have a solution with exactly one open facility f_i, and the solution exists if and only if f_i is contained within at least r intervals. Otherwise $f_i > t(c_1)$, then $SP(i)$ may have a solution with two or more open facilities. In this case $SP(i)$ has a solution if and only if for some $j < i$, $SP(j)$ has a solution, there is no customer c with $f_j < s(c) \leq t(c) < f_i$, and there are at least r customers in $C \setminus C_j$ containing f_i. Intuitively f_j is a possible second rightmost open facility in a solution of $SP(i)$.

We fix the $SP(j)$ with minimum j, if $SP(i)$ has a solution, and we say f_j the mate of f_i, and denoted as $mate(f_i)$. We have the following lemma.

Lemma 6.14. *If $SP(i)$ and $SP(i+1)$ have solutions, then $mate(f_i) \leq mate(f_{i+1})$.*

Proof. For a contradiction assume that $mate(f_i) > mate(f_{i+1})$. Let $f_j = mate(f_i)$ and $f_{j'} = mate(f_{i+1})$. By Lemma 6.13 we have $t(c_{z(j)}) \geq t(c_{z(j')})$. Since $f_{j'}$ is mate of f_{i+1}, there is no customer c such that $f_{j'} < s(c) \leq t(c) < f_{i+1}$. If $t(c_{z(j)}) < f_i$, then $f_{j'}$ is also a mate of f_j, a contradiction. Now if $t(c_{z(j)}) \geq f_j$, then $f_{j'}$ is a mate of f_j since $t(c_{z(j')}) \leq t(c_{z(j)})$, a contradiction. □

We now have the following lemma.

Lemma 6.15. *Let f_i be a facility with $f_i > t(c_1)$ and for some $j < i$, $SP(j)$ has a solution, and $C \setminus C_j$ contains no customer c with $f_j < s(c)$ and $t(c) < f_i$. Fix the $SP(j)$ with minimum j. Then the following holds.*
(a) If $C \setminus C_j$ has less than r customers containing f_i, then no facility $f_{j'}$ with $f_{j'} \geq f_j$ is a mate of f_i, and $SP(i)$ has no solution.
(b) If $SP(i+1)$ has a solution, then $mate(f_{i+1}) \geq f_j$.

Proof. (a) By Lemma 6.13 for any facility $f_{j'} \geq f_j$, if $SP(j')$ has a solution, then $t(c_{z(j')}) \geq t(c_{z(j)})$. Thus the number of customers in $C \setminus C_{j'}$ containing f_i in their interval is less than r.
(b) Assume for a contradiction that $mate(f_{i+1}) \leq f_j$. Let $f_{i'} = mate(f_{i+1})$. Thus there is no customer c with $f_{i'} < s(c)$ and $t(c) < f_{i+1}$. Since $f_{i'} \leq f_i \leq f_{i+1}$, there is no customer c such that $f_{i'} < s(c)$ and $t(c) < f_i$. Hence, $f_{i'}$ is the leftmost facility such that $SP(i')$ has a solution and there is no customer c with $f_{i'} < s(c)$ and $t(c) < f_i$, a contradiction. □

By Lemma 6.14 and 6.15, we observe that we can search for $mate(f_{i+1})$ from where the search for mate of $mate(f_i)$ ends. We now give the following Algorithm called **Proper-interval-*r*-gather**.

If the intervals are sorted according to their right end-points and the facilities are ordered from left to right, then we can preprocess the set of customers containing

Algorithm 5 Proper-interval-r-gather(C, F)

if $|C| < r$ or $F = \emptyset$ **then**
  return $\emptyset$
end if
$i \leftarrow 1$
/* One open facility */
while $f_i \leq t(c_1)$ **do**
  if $f_i \geq s(c_r)$ **then**
    $z(i) \leftarrow r$
    $mate(i) \leftarrow -1$ /* No mate */
  end if
  $i \leftarrow i + 1$
end while
$j \leftarrow 1$
/* Two or more open facilities */
while $i \leq m$ **do**
  $C_i \leftarrow \{c_1, c_2, \cdots, c_{z(i)}\}$
  while $j \leq i$ **do**
    if $C \backslash C_j$ has at least r customers containing f_i and $C \backslash C_j$ has no customer c with $f_j < s(c)$ and $t(c) < f_i$ **then**
      /* $SP(i)$ has a solution */
      $z(i) \leftarrow$ index of the r-th customer in $C \setminus C_j$ containing f_i
      $mate(i) \leftarrow j$
      break
    end if
    if There is no customer between f_j and f_i, and $C \setminus C_j$ has less than r customers containing f_i **then**
      break /* $SP(i)$ has no solution, Lemma 6.15(a) */
    end if
    $j \leftarrow j + 1$
  end while
  $i \leftarrow i + 1$
end while
if Some $SP(i)$ with $f_i \geq s(c_n)$ has a solution **then**
  $j, last, A \leftarrow mate(i), n,$ empty assignment
  while $j \neq -1$ **do**
    $A' \leftarrow$ assign $\{c_{z(j)+1}, c_{z(j)+2}, \ldots, c_{last}\}$ to f_i
    Add assignment A' to A
    $last, i \leftarrow z(j), j$
    $j \leftarrow mate(i)$

38: **end while**
39: **return** A
40: **end if**
41: **return** $\emptyset$

each facility in linear time. Each customer and each facility have to be processed for a constant number of times. Hence the algorithm runs in $O(n+m)$ time. We thus have the following theorem.

Theorem 6.16. *Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities on a line and $C = \{c_1, c_2, \cdots, c_n\}$ be a set of customers where each customer c_i has an interval $I_i = [s(c_i), t(c_i)]$ and no interval is contained within any other interval. The algorithm **Proper-interval-r-gather** decides whether there is an interval r-gathering of C to F, and constructs one if exists in $O(n+m)$ time.*

We now give an outline of the algorithm to solve uncertain r-gathering problem on a line where the customer locations are specified by well-separated uniform distributions. We first explain about preprocessing. Computing the function $E[d(p, c_i)]$ for all the customers takes $O(n)$ time. We can compute the expected distances between customer c_i and all the facilities in $O(m)$ time. Since the function $E[d(p, c_i)]$ is unimodal, the expected distances between c_i and all the facilities can be sorted in $O(m)$ time. Computing the expected distances between each pair of customers and facilities takes $O(mn)$ time and we can merge the of n sorted list of expected distances in $O(mn \log n)$ time using min-heap. Thus we need $O(mn \log n)$ time for the preprocessing. Now we explain about the main algorithm. We do binary search on the ordered list of expected distances to find the optimal r-gathering. Given b we can compute the (c, b)-intervals for all customers in $O(n)$ time. The (c, b)-intervals can be sorted in $O(n \log n)$ time. Then solving each decision instance takes $O(m+n)$ time. Thus we need $O(n \log n + m)$ time to solve the decision instance. To find the optimal solution by binary search we need to solve the decision instances $\log mn$ times, so $O((n \log n + m + n) \log mn)$ in total. Hence the running time is $O(mn \log n + (n \log n + m) \log mn)$. Thus we have the following theorem.

Theorem 6.17. *Let $F = \{f_1, f_2, \cdots, f_m\}$ be a set of facilities on a line and $C = \{c_1, c_2, \cdots, c_n\}$ be a set of customers where each customer c_i has a well-separated uniform distribution. Then an optimal r-gathering of C to F can be constructed in $O(mn \log n + (n \log n + m) \log mn)$ time.*

7. Conclusion

In this paper we presented an $O(n + d^d r^d dr \log d)$ time algorithm to solve rhe r-gather clustering problem when all customers are lying on a star with d rays. We also gave an $O(n+m+(d+\log m)d^4r^2+d^dr^d2^ddr\log d)$ time algorithm to solve the r-gathering problem when all customers and facilities are lying on a star with d rays.

We also showed the hardness of min-max-sum r-gathering problem on a star. We also give an $O(nk+mn\log n+(m+n\log kn+nr^{\frac{n}{r}})\log mn)$ time algorithm for the one-dimensional r-gathering problem when the customer locations are given by piecewise uniform functions of at most $k+1$ pieces, and an $O(mn\log n+(n\log n+m)\log mn)$ time algorithm for the one-dimensional r-gathering problem when the customer locations are given by well-separated uniform distributions.

Acknowledgments

We thank CodeCrafters International and Investortools, Inc. for supporting the first author under the grant "CodeCrafters-Investortools Research Grant".

References

[1] P. K. Agarwal, S. Cheng, Y. Tao, and K. Yi. Indexing uncertain data. In *Proceedings of the Twenty-Eigth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS 2009*, pages 137–146, 2009.

[2] P. K. Agarwal, A. Efrat, S. Sankararaman, and W. Zhang. Nearest-neighbor searching under uncertainty I. *Discrete & Computational Geometry*, 58(3):705–745, 2017.

[3] P. K. Agarwal, S. Har-Peled, S. Suri, H. Yildiz, and W. Zhang. Convex hulls under uncertainty. *Algorithmica*, 79(2):340–367, 2017.

[4] G. Aggarwal, R. Panigrahy, T. Feder, D. Thomas, K. Kenthapadi, S. Khuller, and A. Zhu. Achieving anonymity via clustering. *ACM Transactions on Algorithms*, 6(3):49:1–49:19, 2010.

[5] S. Ahmed, S. Nakano, and M. S. Rahman. One-dimensional r-gathering under uncertainty. In *Proceedings of Algorithmic Aspects in Information and Management*, volume 11640 of Lecture Notes in Computer Science, pages 31–42. Springer Nature Switzerland, 2019.

[6] S. Ahmed, S. Nakano, and M. S. Rahman. r-gatherings on a star. In *Proceedings of the 13th International Workshop on Algorithms and Computation*, volume 11355 of Lecture Notes in Computer Science, pages 31–42. Springer Nature Switzerland, 2019.

[7] T. Akagi and S. Nakano. On r-gatherings on the line. In *Proceedings of Frontiers in Algorithmics*, volume 9130 of Lecture Notes in Computer Science, pages 25–32, Cham, 2015. Springer International Publishing.

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

[9] Z. Drezner and H. W. Hamacher. *Facility Location: Applications and Theory*. Springer, New York, 2004.

[10] M. R. Garey and D. S. Johnson. *Computers and Intractability; A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, NY, USA, 1990.

[11] S. Guha, A. Meyerson, and K. Munagala. Hierarchical placement and network design problems. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 603–612, 2000.

[12] Y. Han and S. Nakano. On r-gatherings on the line. In *Proceedings of FCS 2016*, pages 99–104, 2016.

[13] P. Kamousi, T. M. Chan, and S. Suri. Closest pair and the post office problem for stochastic points. *Computational Geometry*, 47(2):214–223, 2014.

[14] D. R. Karget and M. Minkoff. Building steiner trees with incomplete global knowledge. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 613–623, 2000.

[15] S. Nakano. A simple algorithm for r-gatherings on the line. In *Proceedings of WALCOM: Algorithms and Computation*, volume 10755 of Lecture Notes in Computer Science, pages 1–7, Cham, 2018. Springer International Publishing.

[16] A. Sarker, W. Sung, and M. S. Rahman. A linear time algorithm for the r-gathering problem on the line (extended abstract). In *Proceedings of WALCOM: Algorithms and Computation*, volume 11355 of Lecture Notes in Computer Science, pages 56–66, Cham, 2019. Springer Nature Switzerland.

[17] L. V. Snyder. Facility location under uncertainty: a review. *IIE Transactions*, 38(7):547–564, 2006.

[18] S. Suri and K. Verbeek. On the most likely voronoi diagram and nearest neighbor searching. *International Journal of Computer Geometry Applications*, 26(3-4):151–166, 2016.

[19] Y. Tao, X. Xiao, and R. Cheng. Range search on multidimensional uncertain data. *ACM Transaction on Database Systems*, 32(3):15, 2007.

[20] H. Wang and J. Zhang. One-dimensional k-center on uncertain data. *Theoretical Computer Science*, 602:114–124, 2015.

[21] M. L. Yiu, N. Mamoulis, X. Dai, Y. Tao, and M. Vaitis. Efficient evaluation of probabilistic advanced spatial queries on existentially uncertain data. *IEEE Transaction on Knowledge and Data Engineering*, 21(1):108–122, 2009.

[22] J. Zeng, G. Telang, M. P. Johnson, R. Sarkar, J. Gao, E. M. Arkin, and J. S. B. Mitchell. Mobile r-gather: Distributed and geographic clustering for location anonymity. In *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, Mobihoc '17, pages 7:1–7:10. ACM, 2017.