On the Complexity of Recursive Tree-Based Algorithms for Computing Distance Histograms

Yi-Cheng Tu [1] and Shaoping Chen [2]

[1]*Department of Computer Science and Engineering, University of South Florida*
4202 E. Fowler Ave., ENB118, Tampa, FL 33620, U.S.A.
ytu@cse.usf.edu

[2]*Department of Mathematics, Wuhan University of Technology*
122 Luosi Road, Wuhan, Hubei, 430070, P. R. China
chensp@whut.edu.cn

***Abstract*— Particle simulation has become a popular research tool in many scientific and engineering fields. Analysis of particle simulation data involves computing functions of all particle-to-particle interactions. One such analytics, the spatial distance histogram (SDH), is of vital importance to scientific discovery from particle simulation data. Algorithms for efficiently SDH processing in large-scale simulation have been proposed in a couple of recent papers. These algorithms all adopt a recursive tree-visiting strategy to process particle distances in the visited tree nodes in batches, thus require less time as compared to the brute-force approach where all pairwise distances have to be computed. The complexity of such algorithms have not been thoroughly studied. In this paper, we present an analysis of such algorithms based on a geometric modeling approach. The main technique is to transform the analysis of particle counts into a problem of quantifying the area of regions where particle distances can be processed in batches by the algorithm. From the analysis, we conclude that the number of particle distances that are left to be processed decreases exponentially with more levels of the tree visited. This leads to a time complexity lower than the quadratic time needed for the brute-force algorithm. Our model is also general in that it works for a wide range of space partitioning options in building the tree.**

I. Introduction

The development of advanced experimental devices and computer simulations have given rise to explosive rendering of data in almost all scientific fields. As a result, scientific data management has gained much momentum in research within the database community in recent years. In addition to the challenges of data storage/retrieval imposed by the gigantic volume of scientific data, we also face the issue of designing efficient algorithms for data querying and analysis. Scientific data analysis often require computation of mathematical (statistical) functions whose complexity goes beyond simple aggregates, which are the only analytics supported by modern DBMSs. In this paper, we are interested in the processing of one type of such queries against *particle simulation data*. With applications in various science [1], [2], [3] and engineering [4] fields, particle simulation (PS) is a type of computer simulation that treats system components (e.g., atoms, molecules, stars, galaxies) as classical entities (i.e., particles) that interact via empirical forces. In PS, one query called the *Spatial Distance Histogram (SDH)* is of vital importance.

The SDH problem can be formally stated as: given the coordinates of N particles in a metric space, draw a histogram that represents the distribution of the pairwise distances between the N points. The histogram has a single parameter l, which is the total number of buckets. Since the dataset is always generated from a simulated system with fixed dimensions, the maximum distance between any two points L_{max} is also fixed. The width of the buckets (i.e., histogram resolution) $p = \frac{L_{max}}{l}$ is often used as the parameter of the query instead. In other words, SDH asks for the counts of pairwise distances that fall into ranges $[0, p), [p, 2p), \cdots, [lp-p, lp)$, repsectively. SDH is needed for computing many critical high-level analytics such as pressure, energy, [1] and structure factor [5] in the simulated systems.

While a naive way to compute SDH takes $O(N^2)$ time, more efficient algorithms have been developed [6], [7]. The main idea of this type of algorithm is to derive the histogram by studying the distance between clusters of points instead of those between two individual points. Although different implementations exist in [6] and [7], both can be abstracted into a recursive tree-based algorithm described in Section II.

This paper presents the complexity analysis of the above algorithm. An important contribution of this paper is the model we develop to accomplish such analysis. The main idea is to transform the number of particles into geometric regions whose area can be represented by closed-form formulae, which make rigorous analysis possible. In addition to the complexity analysis of the exact SDH algorithm, our model also builds the foundation for an approximate algorithm with time complexity depending only on a controlled error bound [6]. The latter provides a more practical solution to process SDH in very large simulation datasets.

II. The algorithm

The algorithm first divides the simulated space into a grid, each cell of which records the number of particles in it. We call such a grid a *density map* and density maps with different cell sizes have to be maintained. For this, the algorithm is named

S. Chen is currently a visiting professor in the Department of Computer Science and Engineering at the University of South Florida (USF). His email at USF is: schen11@cse.usf.edu

```
Procedure RESOLVETWOCELLS (A, B)
1  if A and B are resolvable
2      add n1n2 to the corresponding bucket
3  else if A and B are not leaf nodes
4      for each child a of A
5          for each child b of B
6              RESOLVETWOCELLS (a, b)
7  else
8      compute all pairwise distances between A and B
       and add each pair to corresponding bucket
```

Fig. 1. Procedure ResolveTwoCells - core of the DM-SDH algorithm.

density map-based SDH (DM-SDH) algorithm. In practice, we organize all particle coordinates into a point region Quad-tree [8] with each node representing a cell (square for 2D data and cube for 3D) in the simulated space. Particle counts of each cell are cached in the corresponding tree node. The total number of levels of the tree is determined in a way such that, on average, there are a small number β of particles in each leaf node. The process of tree construction can be accomplished in $O(N)$ time.

The focal point of this algorithm is a procedure named RESOLVETWOCELLS (Fig. 1). To resolve two cells $\mathbf{A}$ and $\mathbf{B}$ (with particle count n_1 and n_2, respectively), we first read the coordinates of the two cells and compute the range of distances between any pair of points, one from $\mathbf{A}$ and one from $\mathbf{B}$. Note that this distance range can be computed in constant time. If this range is contained in the range of a histogram bucket i, we say $\mathbf{A}$ and $\mathbf{B}$ are *resolvable* and they *resolve into* bucket i. In this case, we simply increment the count of bucket i by n_1n_2. If the two cells are not resolvable, we recursively make attempts to resolve all their children. In case we have reached the lowest level of the tree, we have to calculate all distances of the particles in the unresolved cells.

The algorithm starts from a certain level of the tree where the diagonal of the cells is no greater than p (otherwise we cannot resolve anything). We denote this level as density map DM_a. First, all intra-cell pairwise distances can be put into the first bucket (with range $[0, p)$). In the next step, RESOLVETWOCELLS is called for all pairs of cells on DM_a.

In the remainder of this paper, we first give a detailed introduction to our geometric model in Section III. In this model, we focus on a simple scenario where the data is of 2D. In building the tree, space partitioning in a node is done by dividing each dimension into two equal sized segments (i.e., regular Quad tree for 2D and Oct tree for 3D). We will then extend our analysis to more generalized conditions on space partitioning and 3D data (Section IV). The complexity analysis based on the model is presented in Section V. We conclude the paper by Section VI.

III. MODELING NUMBER OF RESOLVABLE PARTICLES

An essential problem our analysis needs to answer is: given a cell $\mathbf{A}$ on the first density map DM_a, how many particles

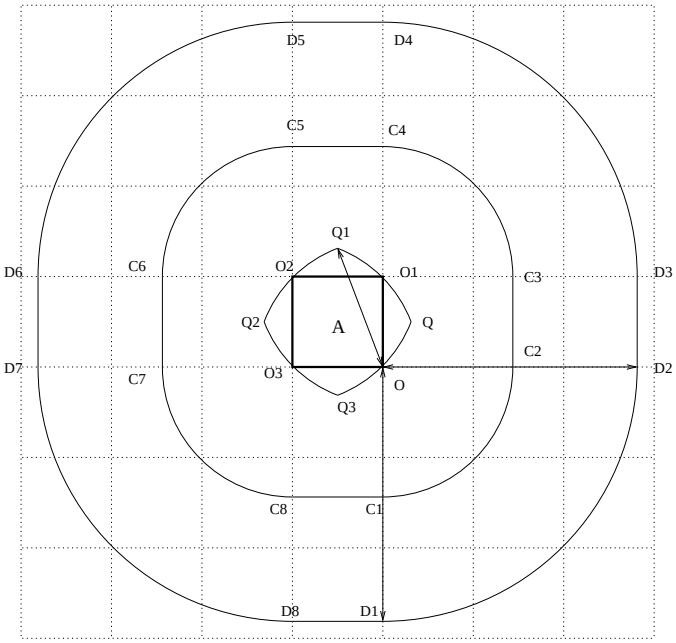


Fig. 2. Boundaries of bucket 1 and bucket 2 regions of cell $\mathbf{A}$, with the bucket width p being exactly $\sqrt{2}\delta$. Here we show arcs $\widehat{Q_1Q_2}$, $\widehat{C_1C_2}$, and $\widehat{D_1D_2}$, all of which are centered at point O.

are contained by those resolvable cells related to $\mathbf{A}$ as we visit more and more levels of density maps? Although this has something to do with the spatial distribution of the particles, we start by analyzing *how much area are covered by the resolvable cells* to simplify the process. To achieve this, we first need to define a theoretical region in which a particle can have distance (to a point in $\mathbf{A}$) that falls into a specific bucket i. We call this region the *bucket i region* of cell $\mathbf{A}$.

A. Basics of the model

In Fig. 2, a cell $\mathbf{A}$ is drawn with four corner points O, O_1, O_2, and O_3. The side length of $\mathbf{A}$ is exactly $\delta = \frac{p}{\sqrt{2}}$. The bucket 1 region of $\mathbf{A}$ is bounded by a curve connected by points C_1 to C_8. This region is drawn as follows: C_1C_2, C_3C_4, C_5C_6, and C_7C_8 are all arcs of 90 degrees centered at the four corners of cell $\mathbf{A}$ and their radii are p; C_2C_3, C_4C_5, C_6C_7, and C_8C_1 are line segments. Note that this is a theoretical "maximum" region where a point can resolve with any point in $\mathbf{A}$. It is easy to see that the area of this region is $\pi p^2 + 4p\delta + \delta^2$. Let us continue to consider distances that fall into the second bucket (i.e., [p, 2p)). Again, the bucket 2 region of $\mathbf{A}$ is of similar shape to the bucket 1 region except the radii of the arcs are $2p$, as drawn in Fig. 2 with a curve connected by points D_1 to D_8. However, points that are too close to $\mathbf{A}$ can only resolve into bucket 1 since their distances to any point in $\mathbf{A}$ will always be smaller than p. These points are contained in a region as follows: on each corner point of $\mathbf{A}$, we draw an arc with radius p on the opposite corner (i.e., arcs QQ_1, Q_1Q_2, Q_2Q_3, and Q_3Q_4). Therefore, the bucket 2 region should not include this

$$g(i) = \begin{cases} (2\pi + 4\sqrt{2} + 1)\delta^2 & i = 1 \\ \left[2\pi i^2 + 4\sqrt{2}i - (i-1)^2\left(8\arctan\sqrt{8(i-1)^2-1} - 2\pi\right) + \sqrt{8(i-1)^2-1}\right]\delta^2 & i > 1 \end{cases} \quad (1)$$

inner region (denoted as region $\mathbf{B}$ hereafter). A more detailed illustration of region $\mathbf{B}$ is shown in Fig. 3.

The area of the bucket 2 region is $\pi(2p)^2 + 8p\delta$ less the area of region $\mathbf{B}$, which consists of eight identical smaller regions such as $\widehat{Q_1O_2}D$ and cell $\mathbf{A}$ itself (Fig. 3). To get the area of $\widehat{Q_1O_2}D$, we first need to know the magnitude of the angle $\angle Q_1OO_2$, which can be computed by

$$\begin{aligned} \angle Q_1OO_2 &= \angle Q_1OE - \angle COE \\ &= \arctan\frac{Q_1E}{EO} - \frac{\pi}{4} \\ &= \arctan\frac{\sqrt{p^2 - \left(\frac{\delta}{2}\right)^2}}{\frac{\delta}{2}} - \frac{\pi}{4} \end{aligned}$$

Thus, the area of sector $\widehat{Q_1O_2}O$ is $\frac{1}{2}p^2\angle Q_1OO_2$. The area of region $\widehat{Q_1O_2}D$ can be obtained by the area of this sector less the area of triangles O_2DC and Q_1CO as follows:

$$\begin{aligned} S_{\widehat{Q_1O_2}D} &= S_{\widehat{Q_1O_2}O} - S_{\triangle O_2DC} - S_{\triangle Q_1CO} \\ &= \frac{1}{2}p^2\left[\arctan\frac{\sqrt{p^2 - \left(\frac{\delta}{2}\right)^2}}{\frac{\delta}{2}} - \frac{\pi}{4}\right] - \frac{1}{2}\left(\frac{\delta}{2}\right)^2 \\ &\quad - \frac{1}{2}\left[\sqrt{p^2 - \left(\frac{\delta}{2}\right)^2} - \frac{\delta}{2}\right]\frac{\delta}{2} \\ &= \frac{1}{2}p^2\left[\arctan\frac{\sqrt{p^2 - \left(\frac{\delta}{2}\right)^2}}{\frac{\delta}{2}} - \frac{\pi}{4}\right] \\ &\quad - \frac{\delta}{4}\sqrt{p^2 - \left(\frac{\delta}{2}\right)^2} \end{aligned}$$

and we have $\pi(2p)^2 + 8p\delta - 8S_{\widehat{Q_1O_2}D} - S_{\mathbf{A}}$ as the area of the bucket 2 region.

The approach to obtain the area of bucket i ($i > 2$) regions is the same as above. For the area of the region formed by the outer boundary, we only need to consider that the arcs in Fig. 3 are of radii ip. The development of a general formula for the area of region $\mathbf{B}$ is trickier. Our efforts lead to the following formula to quantify the area of the bucket i region:

$$g(i) = \begin{cases} \pi p^2 + 4p\delta + \delta^2 & i = 1 \\ \pi(ip)^2 + 4ip\delta - [8A(i) + B(i)\delta^2] & i \geq 2 \end{cases}$$

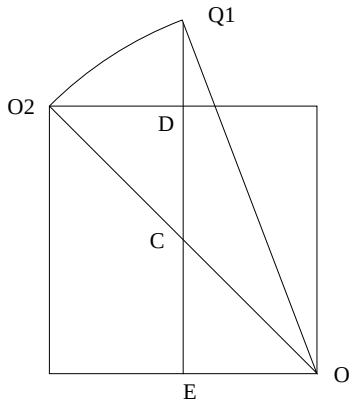


Fig. 3. An illustration on how to compute the area of region $\mathbf{B}$ (i.e., $QQ_1Q_2Q_3$ formed by four arcs in Fig. 2). Here we only show arc Q_1O_2, which is a half of one of the arcs Q_1Q_2.

where $B(i) = [2(i-1)-1]^2 - 1$ and

$$\begin{aligned} A(i) = \frac{1}{2}[(i-1)p]^2 &\left[\arctan\frac{\sqrt{[(i-1)p]^2 - \left(\frac{\delta}{2}\right)^2}}{\frac{\delta}{2}} - \frac{\pi}{4}\right] \\ &- \frac{1}{2}\frac{\delta}{2}\left[\sqrt{[(i-1)p]^2 - \left(\frac{\delta}{2}\right)^2} - \frac{\delta}{2}\right] \\ &- \frac{1}{2}\left[(i-2)\delta + \frac{\delta}{2}\right]^2 \end{aligned}$$

Since we have $p = \sqrt{2}\delta$, the above equation becomes Eq. (1) shown on top of this page.

B. Coverable regions

Eq. (1) gives the area of a theoretical region that contains all particles that could possibly resolve into a given bucket with a point in cell $\mathbf{A}$. Now let us study how much of this region can be resolved in our algorithm under different levels of density maps. We call the region that consists of all resolvable cells the *coverable region*.

1) Case 1: the first bucket: Let us start our discussions on the situation of bucket 1. In Fig. 4, we show the coverable regions of three different density map levels: $m = 1$, $m = 2$, and $m = 3$, as represented by blue-colored lines and denoted as $\mathbf{A}'$ in all subgraphs. For $m = 1$, the resolvable cells are only those surrounding $\mathbf{A}$. All other cells, even those entirely contained by the bucket 1 region, do not resolve with any level 1 subcell of $\mathbf{A}$. As we increase m, the region $\mathbf{A}'$ grows in area, with its boundary approaching that of the bucket 1 region. To represent the area of $\mathbf{A}'$, we need to develop a continuous line to approximate its boundary. One critical observation here is: the furtherest cells in $\mathbf{A}'$ are those that can resolve with cells

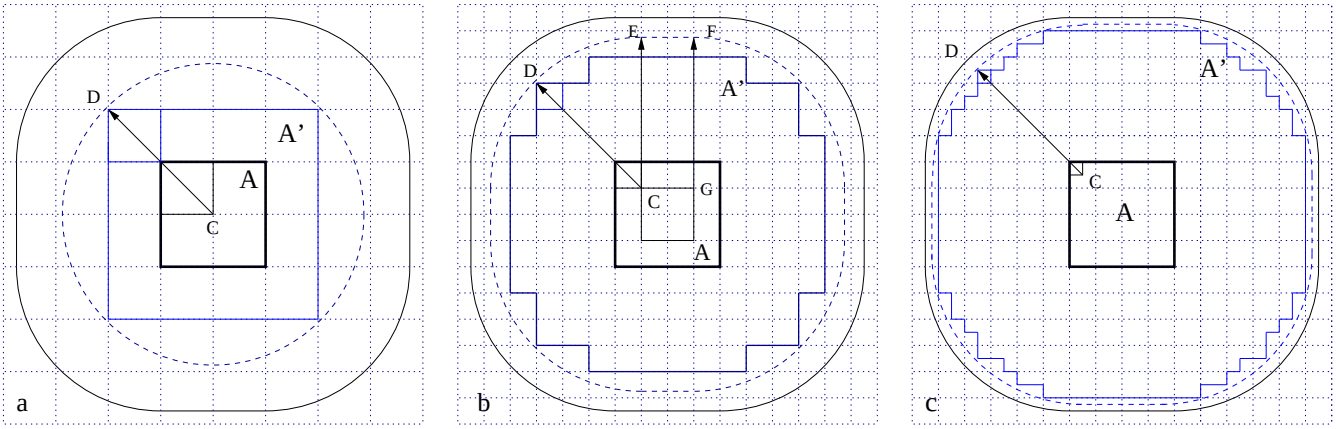


Fig. 4. Actual (solid blue line) and approximated (dotted blue line) coverable regions for bucket 1 under: a. $m = 1$; b. $m = 2$; and c. $m = 3$. Outer solid black lines represent the theoretical bucket 1 region. All arrowed line segments are drawn from the centers to the corresponding arcs with radius p.

on the outer rim of $\mathbf{A}$. For example, the cell cornered at point D resolves with the cell cornered at point C in $\mathbf{A}$. If we draw a 90-degree arc centered at C, the arc goes through D and all cells on the northwestern corner of $\mathbf{A}'$ are bounded by this arc. To approximate the boundary of $\mathbf{A}'$, we can draw such an arc at all four corners of the graph and connect them with line segments (e.g., EF connecting the northwestern and northeastern arcs centered at point G in Fig. 4b), as shown by the blue dotted line. Obviously, this line approaches the theoretical boundary as m increases because the center of the arcs (e.g., point C) move further to the corner points of $\mathbf{A}$ as the cells become smaller. Note this line gives rise to an optimistic approximation of $\mathbf{A}'$. In a moment, we will show that this overestimation will not harm our analysis on the running time of DM-SDH. The area of the coverable region for bucket 1 at level m can be expressed as

$$S_{\mathbf{A}'} = \pi p^2 + 4p\left(\delta - \frac{2\delta}{2^m}\right) + \left(\delta - \frac{2\delta}{2^m}\right)^2 \qquad (2)$$

where the first item πp^2 is the area of the four 90-degree sectors centered at point C, the second item is the area of the four rectangles (e.g., EFGC in Fig. 4b) connecting the four sectors. We also need to add the area of the small square (with side CG in Fig. 4b) within cell $\mathbf{A}$, which is given by the last item.

2) Case 2: the second bucket and beyond: The cases of buckets beyond the first one are more complicated. First of all, the outer boundary of the bucket i ($i \geq 2$) regions can be approximated using the same techniques we introduced for bucket 1 (Section III-B.1). Therefore, we can use the following generalized form of Eq. (2) to quantify the area of the region formed by the outer boundaries only.

$$S_{out}(i) = \pi(ip)^2 + 4ip\left(\delta - \frac{2\delta}{2^m}\right) + \left(\delta - \frac{2\delta}{2^m}\right)^2 \qquad (3)$$

However, we also need to disregard the cells that lie in the inner boundary (e.g., those within or near region $\mathbf{B}$). To quantify the area of the region contained by the inner boundary, we need to consider the cases of $m = 1$ and $m > 1$ separately.

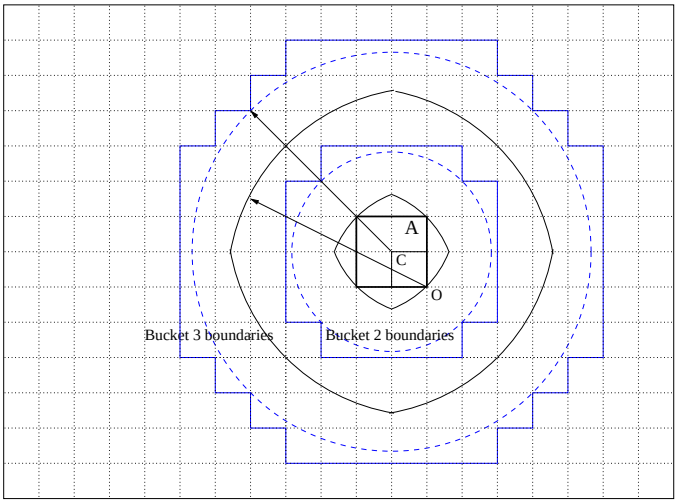


Fig. 5. Inner boundaries of the coverable regions of buckets 2 and 3 under $m = 1$. All arrowed line segments are of length $2p$.

Let us first study the case of $m = 1$. Fig. 5 shows examples with $m = 1$ with respect to the second and the third bucket. It is easy to see that any cell that contains a segment of the theoretical region $\mathbf{B}$ boundary will not resolve into bucket i because they can only resolve into bucket $i - 1$. Furthermore, there are more cells that resolve into neither bucket $i - 1$ nor bucket i. Here our task is to find a boundary to separate those $m = 1$ cells that can resolve into bucket i with any subcell in $\mathbf{A}$ and those that cannot. Such boundaries for buckets 2 and 3 are shown in Fig. 5 as solid blue lines. The boundary can be generated as follows: on each quadrant (e.g., northwest) of cell $\mathbf{A}$, we draw an arc (dotted blue line) centered at the corner point C of the furthest (e.g., southeast) subcell of $\mathbf{A}$ with radius $(i - 1)p$. Any cell that contains a segment of this arc cannot resolve into bucket i (because they are too close to $\mathbf{A}$) but the cells beyond this line can. Therefore, we can also

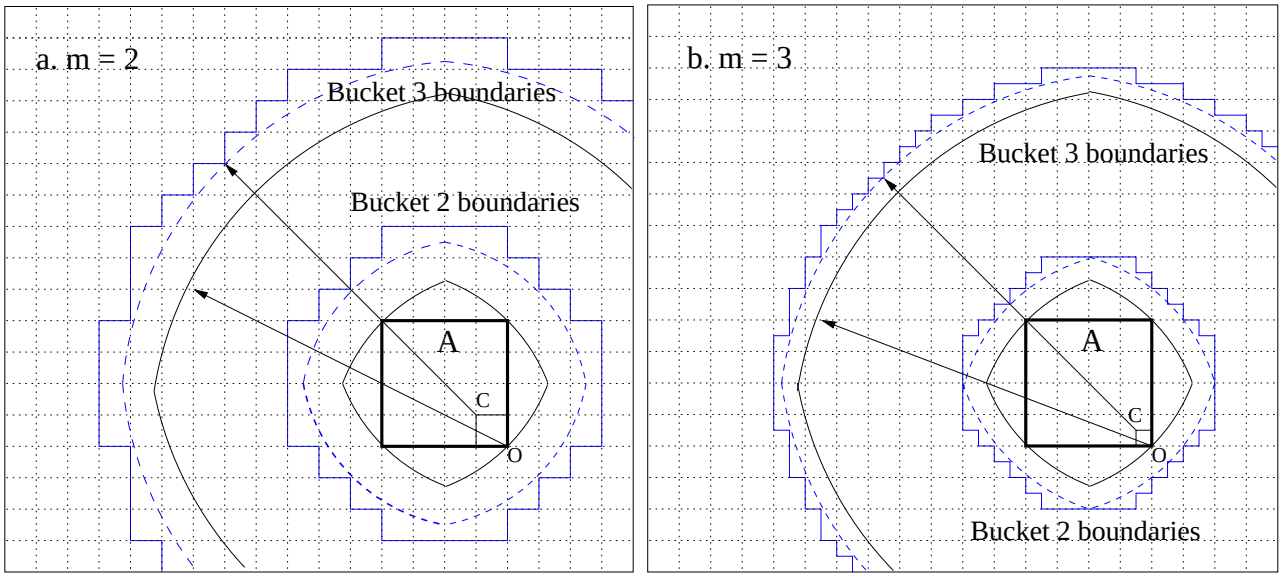


Fig. 6. Inner boundaries of the coverable regions of buckets 2 and 3 under $m = 2$ and $m = 3$. All arrowed line segments are of length $2p$.

use these arcs to approximate the zigzagged real boundaries. Let us denote the region bounded by this approximate curve as region $\mathbf{B}'$. For $m = 1$, the arcs on all four quadrants share the same center C therefore they form a circle as region $\mathbf{B}'$. The radii of the circles are exactly $(i-1)p$ for bucket i. Note that this, again, could give rise to an optimistic approximation of the area of coverable regions. Therefore, the area of the coverable region for $m = 1$ and $i \geq 2$ is:

$$S_{\mathbf{A}'} = \pi(ip)^2 - \pi[(i-1)p]^2 \tag{4}$$

where the first item is the area of the region formed by the approximated outer boundary, which is given as a special case of Eq. (3) for $m = 1$ and happens to be a circle; and the second item is that of the region formed by the approximated inner boundary (i.e., region $\mathbf{B}'$).

For the case of $m > 1$, we can use the same technique described for the case of $m = 1$ to generate the curves to form region $\mathbf{B}'$. However, these curves are no longer a series of circles. In Fig. 6, we can find such curves for buckets 2 and 3 under m values of 2 and 3. As the four arcs on different quadrants no longer share the same center, the region $\mathbf{B}'$ boundaries (dotted blue lines) are of similar shapes to the theoretical region $\mathbf{B}$ boundaries (solid black lines). From the graphs, it is easy to see that the approximated curve fits the actual boundary better as m increases. Here we skip the formal proof as it is straightforward. Furthermore, it also converges to the region $\mathbf{B}$ boundary when m gets bigger. This is because the centers of the two arcs (with the same radii), points C and O, become closer and closer when the cell size decreases (as a result of the increase of m).

The area of region $\mathbf{B}'$ can be computed in the same way as that of region $\mathbf{B}$, with the help of an illustration in Fig. 7.

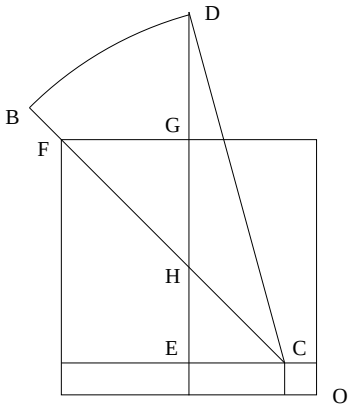


Fig. 7. An illustration on how to compute the area of region formed by four arcs in Fig. 6. Here we only show half of one of the arcs.

First, we get the magnitude of angle BCD by

$$\begin{aligned}\angle BCD &= \angle DCE - \angle FCE \\ &= \arctan\frac{DE}{EC} - \frac{\pi}{4} \\ &= \arctan\frac{\sqrt{[(i-1)p]^2 - \left(\frac{\delta}{2} - \frac{\delta}{2^m}\right)^2}}{\frac{\delta}{2} - \frac{\delta}{2^m}} - \frac{\pi}{4}\end{aligned}$$

From now on, let us define θ as a function of m for the convenience in further discussions:

$$\theta_m = \frac{1}{2} - \frac{1}{2^m}.$$

The area of the sector $\widehat{BDC}$ is $\frac{1}{2}[(i-1)p]^2\angle BCD$, and the

$$f(i,m) = \begin{cases} \left[2\pi + 4\sqrt{2} + 1 - (8\sqrt{2}+4)\frac{1}{2^m} + \frac{4}{2^{2m}}\right]\delta^2 & i = 1, m \geq 1 \\ \left[2\pi(2i-1)\right]\delta^2 & i > 1, m = 1 \\ \left\{2\pi i^2 + 4\sqrt{2}i - (8\sqrt{2}i+4)\frac{1}{2^m} + \frac{4}{2^{2m}} - 8\left[\begin{array}{c}(i-1)^2\left(\arctan\frac{\gamma_m}{\theta_m} - \frac{\pi}{4}\right) \\ -\frac{1}{2}\theta_m(\gamma_m - \theta_m)\end{array}\right] + 1\right\}\delta^2 & i > 1, m > 1 \end{cases} \quad (5)$$

area of the region $\widehat{BDGF}$ is

$$\begin{aligned} S_{\widehat{BDGF}} &= S_{\widehat{BDC}} - S_{\triangle DHC} - S_{\triangle FGH} \\ &= \frac{1}{2}[(i-1)p]^2 \angle BCD - \frac{1}{2}EC(DE - HE) - \frac{\delta^2}{8} \\ &= \frac{1}{2}[(i-1)p]^2\left[\arctan\frac{\sqrt{[(i-1)p]^2 - \delta^2\theta_m^2}}{\delta\theta_m} - \frac{\pi}{4}\right] \\ &\quad - \frac{\delta}{2}\theta_m\left[\sqrt{[(i-1)p]^2 - (\delta\theta_m)^2} - \delta\theta_m\right] - \frac{\delta^2}{8} \end{aligned}$$

Finally, we get the area of the coverable region for $i \geq 2, m > 1$ as

$$\begin{aligned} S_{A'} &= S_{out}(i) - 8S_{\widehat{BDGF}} - S_A \\ &= \pi(ip)^2 + 4ip\left(\delta - \frac{2\delta}{2^m}\right) + \left(\delta - \frac{2\delta}{2^m}\right)^2 \\ &\quad - 4[(i-1)p]^2\left[\arctan\frac{\sqrt{[(i-1)p]^2 - \delta^2\theta_m^2}}{\delta\theta_m} - \frac{\pi}{4}\right] \\ &\quad + 4\delta\theta_m\left[\sqrt{[(i-1)p]^2 - (\delta\theta_m)^2} - \delta\theta_m\right] \end{aligned} \quad (6)$$

In summary, let us denote the area of the coverable region $\mathbf{A}'$ for bucket i under different m values as $f(i,m)$. By combining and simplifying Equations (2), (4), and (6) with $p = \sqrt{2}\delta$, we get Eq. (5) shown on top of this page, in which $\gamma_m = \sqrt{2(i-1)^2 - \theta_m^2}$.

C. Covering factor

In this section, we give a quantitative analysis on the relationship between $f(i,m)$ and the area of the theoretical region $g(i)$ for all buckets. For that purpose, given any density map level m, we define the *covering factor* $c(m)$ as the ratio of the total area of the coverable regions to that of the theoretical bucket i regions for all i. However, the quantity that is more related to our analysis is the *non-covering factor* $\alpha(m) = 1 - c(m)$. Specifically, we have

$$\alpha(m) = \frac{\sum_{i=1}^{l}[g(i) - f(i,m)]}{\sum_{i=1}^{l} g(i)} \quad (7)$$

The quantity $\alpha(m)$ is important in that it can directly tell how many cell pairs are resolvable on a given density map level (as the total number of cell pairs is always known for each level). Before investigating the features of $\alpha(m)$, let us define two relevant quantities, the total area of bucket regions for all buckets G, and that of all coverable regions F. Being summations over all buckets of $g(i)$ and $f(i,m)$, they can be expressed as functions of the total bucket number l. We also remove the common factor δ^2 from both Eq. (1) and Eq. (5) for the convenience of displaying equations. First, we have

$$\begin{aligned} G(l) &= \frac{\sum_{i=1}^{l} g(i)}{\delta^2} \\ &= 1 + \sum_{i=1}^{l}\left(2\pi i^2 + 4\sqrt{2}\right) \\ &\quad - \sum_{i=2}^{l}\left[(i-1)^2(8\arctan\sigma_i - 2\pi) - \sigma_i\right] \\ &= 1 + \frac{2}{3}l\left(3\sqrt{2} + 3\sqrt{2}l + \pi + 2l^2\pi\right) \\ &\quad - \sum_{i=2}^{l}\left[(i-1)^2(8\arctan\sigma_i - 2\pi) - \sigma_i\right] \end{aligned} \quad (8)$$

where $\sigma_i = \sqrt{8(i-1)^2 - 1}$. The area of total coverable regions is considered in two cases. For $m = 1$, we get

$$\begin{aligned} F(l,1) &= \frac{\sum_{i=1}^{l} f(i,1)}{\delta^2} \\ &= 2\pi + 2\pi\sum_{i=2}^{l}(2i-1) = 2\pi l^2 \end{aligned} \quad (9)$$

and for $m > 1$, we have the following formula:

$$\begin{aligned} F(l,m) &= \frac{\sum_{i=1}^{l} f(i,m)}{\delta^2} \\ &= 2^{2-2m} - 2^{2-m} + 1 + 2\sqrt{2}l - 2^{\frac{5}{2}-m}l \\ &\quad + 2\sqrt{2}l^2 - 2^{\frac{5}{2}-m}l^2 + \frac{3}{2}l\pi + \frac{4}{3}l^3\pi \\ &\quad - 8\sum_{i=2}^{l}(i-1)^2\arctan\frac{\sqrt{2(i-1)^2 - \theta_m^2}}{\theta_m} \\ &\quad + 4\sum_{i=2}^{l}\theta_m\sqrt{2(i-1)^2 - \theta_m^2} \end{aligned} \quad (10)$$

With the above definitions, we develop the most important result in our analysis in the following lemma.

Lemma 1: For any given standard SDH query with bucket width p, let DM_a be the first density map the DM-SDH algorithm starts running, and $\alpha(m)$ be the non-covering factor of a density map that lies m levels below DM_a (i.e., map DM_{a+m}). We have

$$\lim_{p\to 0}\frac{\alpha(m+1)}{\alpha(m)} = \frac{1}{2}.$$

Proof: From Eq. (7), we easily get

$$\frac{\alpha(m+1)}{\alpha(m)} = \frac{G(l) - F(l, m+1)}{G(l) - F(l, m)}$$

Plugging Eq. (8), Eq. (9), and Eq. (10) into the above formula, we get $\frac{\alpha(m+1)}{\alpha(m)} = \frac{A(m)}{B(m)}$ where

$$\begin{aligned} A(m) = & \frac{2}{2^m} - \frac{1}{4^m} + \frac{2^{\frac{3}{2}}}{2^m}(l + l^2) + \sum_{i=2}^{l} \sqrt{8(i-1)^2 - 1} \\ & - 4\sum_{i=2}^{l} \theta_{m+1} \sqrt{2(i-1)^2 - \theta_{m+1}^2} \\ & + 8\sum_{i=2}^{l} (i-1)^2 \arctan \frac{\sqrt{8(i-1)^2 - \theta_{m+1}^2}}{\theta_{m+1}} \\ & - 8\sum_{i=2}^{l} (i-1)^2 \arctan \sqrt{8(i-1)^2 - 1} \end{aligned} \tag{11}$$

and

$$\begin{aligned} B(m) = & \frac{4}{2^m} - \frac{4}{4^m} + \frac{2^{\frac{5}{2}}}{2^m}(l + l^2) + \sum_{i=2}^{l} \sqrt{8(i-1)^2 - 1} \\ & - 4\sum_{i=2}^{l} \theta_{m} \sqrt{2(i-1)^2 - \theta_{m}^2} \\ & + 8\sum_{i=2}^{l} (i-1)^2 \arctan \frac{\sqrt{8(i-1)^2 - \theta_{m}^2}}{\theta_{m}} \\ & - 8\sum_{i=2}^{l} (i-1)^2 \arctan \sqrt{8(i-1)^2 - 1}\ , \end{aligned} \tag{12}$$

in which $\theta_m = \frac{1}{2} - \frac{1}{2^m}$ and $\theta_{m+1} = \frac{1}{2} - \frac{1}{2^{m+1}}$.

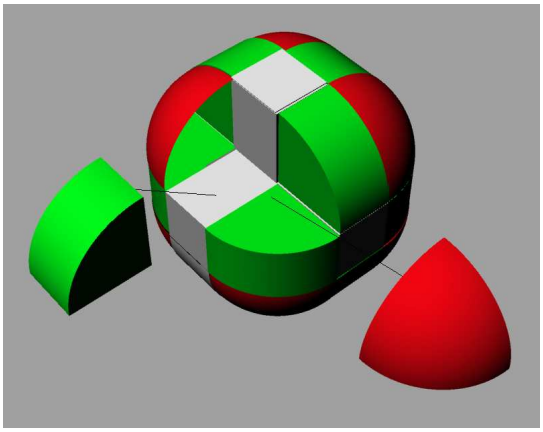
The case of $p \to 0$ is equivalent to $l \to \infty$. Despite their formidable length and complexity, $A(m)$ and $B(m)$ have the following feature

$$\lim_{l \to \infty} \frac{A(m)}{B(m)} = \frac{1}{2} \tag{13}$$

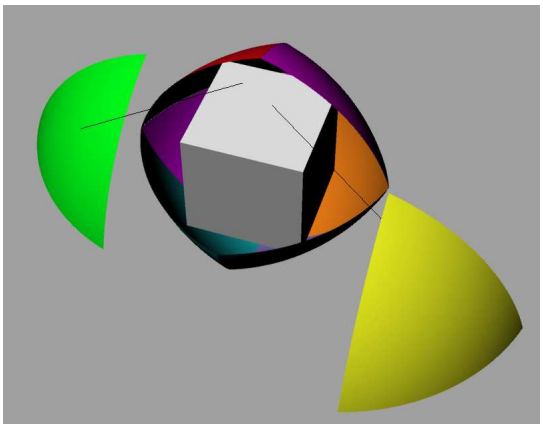
and this concludes the proof. More details on derivation of Eq. (13) can be found in Appendix I. ■

Lemma 1 is important in that it shows the number of non-resolvable cell pairs decreases exponentially (by half) when more levels of the tree are visited. In ResolveTwoCells, if a cell pair is not resolved, we have to make 16 recursive calls to the same routine for the 4 children of each cell. Lemma 1 says that we can expect $16 \times 0.5 = 8$ pairs of the children to be resolvable. This greatly eases our analysis of the time complexity of DM-SDH (Section V).

While shown in the form of a limit under large l (i.e., small p), Lemma 1 also works well under small l values. This can be effectively verified by numerical results due to the closed form of Eq. (11) and Eq. (12). From the results shown in Table I, we can easily see that the ratio of $\alpha(m+1)$ to $\alpha(m)$ quickly converges even when l is very small.



(a) Outer boundary of the bucket 1 region.



(b) Inner boundary of the bucket 2 region.

Fig. 8. Geometric structure of the bucket 1/2 regions for 3D data.

IV. Extensions

A. 3D analysis

The strategies used to achieve the above analysis can be extended to 3D data. The outer and inner boundaries of bucket i regions are illustrated in Fig. 8. The analysis should be based on the volume of relevant regions surrounding a cube $\mathbf{A}$ with side length δ. The bucket 1 region (Fig.8(a)) of $\mathbf{A}$ consists of the following components: 1) quarter cylinders (green) with length δ and radius $p = \sqrt{3}\delta$; 2) one-eighth of a sphere (red) with radius p; 3) cuboids (white) with dimensions δ, δ, and p; and 4) cube $\mathbf{A}$ itself (not shown). There are eight pieces of each of the first two items and six pieces of item 3. The inner boundary (region $\mathbf{B}$) of the bucket 2 region (Fig. 8(b)) consists of eight identical portions of a spherical surface centered at the opposite corner of $\mathbf{A}$ with radius p. Note that the projection of these regions on 2D are exactly those found in Fig. 2. Again, the shape of the region does not change with respect to bucket number i - we only need to change p to ip.

The volume of the bucket i region can thus be expressed as

$$g(i) = \begin{cases} \frac{4}{3}\pi p^3 + 6p\delta^2 + 3\pi p^2\delta + \delta^3, & i = 1 \\ \frac{4}{3}\pi (ip)^3 + 6ip\delta^2 + 3\pi (ip)^2\delta + \delta^3 - v(i,p,\delta), & i > 1 \end{cases}$$

TABLE I

VALUES OF $\alpha(m+1)/\alpha(m)$ OF 2D DATA UNDER DIFFERENT VALUES OF m AND l. COMPUTED WITH MATHEMATICA 6.0. PRECISION UP TO THE 6TH DIGIT AFTER DECIMAL POINT.

Map levels	**Total Number of Histogram Buckets (l)**							
	2	4	8	16	32	64	128	256
m=1	0.508709	0.501837	0.50037	0.50007	0.500012	0.500002	0.5	0.5
m=2	0.503786	0.500685	0.500103	0.500009	0.499998	0.499999	0.499999	0.5
m=3	0. 501749	0.500282	0.500031	0.499998	0.499997	0.499999	0.5	0.5
m=4	0. 500838	0.500126	0.50001	0.499997	0.499998	0.499999	0.5	0.5
m=5	0. 50041	0.500059	0.500004	0.499998	0.499999	0.5	0.5	0.5
m=6	0.500203	0.500029	0.500002	0.499999	0.499999	0.5	0.5	0.5
m=7	0.500101	0.500014	0.500001	0.499999	0.5	0.5	0.5	0.5
m=8	0.50005	0.500007	0.5	0.5	0.5	0.5	0.5	0.5
m=9	0.500012	0.500003	0.5	0.5	0.5	0.5	0.5	0.5
m=10	0.500025	0.500002	0.5	0.5	0.5	0.5	0.5	0.5

TABLE II

VALUES OF $\alpha(m+1)/\alpha(m)$ OF 3D DATA UNDER DIFFERENT VALUES OF m AND l. COMPUTED WITH MATHEMATICA 6.0. PRECISION UP TO THE 6TH DIGIT AFTER DECIMAL POINT.

Map levels	**Total Number of Histogram Buckets (l)**							
	2	4	8	16	32	64	128	256
m=1	0.531078	0.509177	0.502381	0.500598	0.50015	0.500038	0.50001	0.500002
m=2	0.514551	0.504128	0.50102	0.500247	0.50006	0.500013	0.500004	0.5
m=3	0.505114	0.500774	0.500051	0.499987	0.499991	0.501551	0.499996	0.500004
m=4	0.498119	0.497695	0.499076	0.499717	0.499931	0.498428	0.5	0.5
m=5	0.490039	0.49337	0.496703	0.499313	0.499811	0.499966	0.5	0.499983
m=6	0.47651	0.485541	0.49586	0.498521	0.499586	0.499897	0.499931	0.499897
m=7	0.448987	0.469814	0.48972	0.497032	0.499241	0.499793	0.499931	0.500138
m=8	0.38559	0.435172	0.478726	0.494029	0.49848	0.499448	0.499862	0.5

where the first four items in both cases represent the volumes of the four components listed above and $v(i,p,\delta)$ is that for the region formed by half of a spherical surface in Fig. 8(b). With $p=\sqrt{3}\delta$, the above equation becomes

$$g(i)=\begin{cases}\left(4\sqrt{3}\pi+6\sqrt{3}+9\pi+1\right)\delta^3 & i=1\\ \left[4\sqrt{3}\pi i^3+6\sqrt{3}i+9\pi i^2+1-v(i,p)\right]\delta^3 & i>1\end{cases}$$

where $v(i,p,\delta)=16V_{\mathbf{B}}$ and

$$\begin{aligned}V_{\mathbf{B}} &= \iint_{\mathbf{B}} dxdy\int_{\delta/2}^{\sqrt{p^2-x^2-y^2}} dz\\ &= \iint_{\mathbf{B}}\left(\sqrt{p^2-x^2-y^2}-\frac{\delta}{2}\right)dxdy\\ &= \int_a^{\frac{\pi}{4}} d\theta\int_b^c\left(\sqrt{p^2-r^2}-\frac{\delta}{2}\right)rdr\\ &= \int_a^{\frac{\pi}{4}}\left[-\frac{1}{3}(p^2-r^2)^{\frac{3}{2}}-\frac{\delta}{4}r^2\right]\Big|_b^c d\theta\\ &= \int_a^{\frac{\pi}{4}}\left[-\frac{\delta^3}{24}+\frac{1}{3}(p^2-b^2)^{\frac{3}{2}}-\frac{\delta}{4}c^2+\frac{1}{16}\frac{\delta^3}{(\sin\theta)^2}\right]d\theta\end{aligned}$$

where $a=\arctan\dfrac{\frac{\delta}{2}}{\sqrt{p^2-2\left(\frac{\delta}{2}\right)^2}}$, $c=\sqrt{p^2-\left(\frac{\delta}{2}\right)^2}$, and $b=\dfrac{\delta}{2\sin\theta}$.

We continue to develop formulae for the coverable regions $f(i,m)$ and non-covering factor $\alpha(m)$ as we do in Section III-B and Section III-C. These formulae can be found in Appendix II. The complexity of such formulae[1] hinders an analytical conclusion on the convergence of $\alpha(m+1)/\alpha(m)$ towards $\frac{1}{2}$. Fortunately, we are able to compute the numerical values of $\alpha(m+1)/\alpha(m)$ under a wide range of inputs. These results (listed in Table II) clearly show that it does converge to $\frac{1}{2}$.

B. General tiling factor

We use a regular tiling [9] approach to partition the space in building the trees, i.e., the subcells are of the same shape (square/cube) as the parent cell. In the previous analysis, for each node, we evenly cut each dimension by half, leading to 2^d partitions (child nodes) on the next level. However, in general, we could cut each dimension into $s>2$ equal segments, giving rise to s^d equal-sized squares or cubes. In this section, we study how this affects the value of $\alpha(m)$.

First, the theoretical bucket regions given by Eq. (1) are not affected. For the coverable regions, we incorporate the tiling factor s into the same reasoning as what we utilize to obtain Eq. (5). One exception here is the case of $m=1, i\geq 2$: the approximate coverable region does not form a series of circles when $s>2$, therefore Eq. (4) does not hold and this case should be handled in the same way as the case of $m>$

[1]We use Mathematica to solve the integration in Eq. (22) and it ended up an equation that occupies 120 pages!

$$f(i,m,s)=\begin{cases}\left[2\pi+4\sqrt{2}+1-(8\sqrt{2}+4)\frac{1}{s^m}+\frac{4}{s^{2m}}\right]\delta^2 & i=1,m\geq 1\\ \left\{2\pi i^2+4\sqrt{2}i-(8\sqrt{2}i+4)\frac{1}{s^m}+\frac{4}{s^{2m}}-8\left[\begin{array}{c}(i-1)^2\left(\arctan\frac{\gamma'_m}{\theta'_m}-\frac{\pi}{4}\right)\\ -\frac{1}{2}\theta'_m\left(\gamma'_m-\theta'_m\right)\end{array}\right]+1\right\}\delta^2 & i>1,m>1\end{cases} \tag{14}$$

$1, i \geq 2$. Skipping the details, we get an improved version of Eq. (5) for $s > 2$ as Eq. (14), where $\theta'_m = \frac{1}{2} - \frac{1}{s^m}$ and $\gamma'_m = \sqrt{2(i-1)^2 - {\theta'_m}^2}$. With Eq. (14) to describe the coverable regions, we can easily generate new equations for the covering factor as a function of m and s. By studying these functions, we get the following lemma.

Lemma 2: With a tiling factor s ($s \in \mathbb{R}^+$), the non-covering factors have the following property

$$\lim_{p\to 0}\frac{\alpha(m+1,s)}{\alpha(m,s)}=\frac{1}{s}.$$

Proof: See Apprendix III for details. ■

Lemma 2 is obviously a nicely-formatted extension of Lemma 1. As Lemma 1, it is well supported by numerical results even under smaller values of l (details not shown in this paper). In Section V, we will discuss the effects of s on the time complexity of DM-SDH.

V. Time complexity of DM-SDH

In DM-SDH, time spent on the following two operations is dominant: (i) attempts to resolve all pairs of visited nodes (both nodes of the pair are always on the same level of the tree); and (ii) pairwise distance calculation for particles in the unresolved leaf nodes. With Lemma 2, we achieve the following analysis of the time complexity of DM-SDH as a function of the input size N.

Theorem 1: In DM-SDH, the time spent on resolving cells is $\Theta\big(N^{\frac{2d-1}{d}}\big)$ where $d \in \{2,3\}$ is the number of dimensions of the data.

Proof: We derive the complexity by studying how the required time changes with the increase of system size N. Since we keep the average number of particles in a leaf node as a constant β, one more level of tree will be built when N increases to s^dN. Thus, by denoting the time spent on resolving cells as T_c, we need to build a recurrence function of time that relates $T_c(s^dN)$ to $T_c(N)$.

For given bucket width p, the starting level DM_a is fixed in DM-SDH. Assume there are I pairs of cells to be resolved on DM_a. On the next level DM_{a+1}, total number of cell pairs becomes Is^{2d}. According to Lemma 2, only one s-th of the I pairs on DM_a will not be resolved, leaving Is^{2d-1} pairs to resolve on DM_{a+1}. On level DM_{a+2}, this number becomes $Is^{2d-1}\frac{1}{s}s^{2d} = Is^{2(2d-1)}$. Therefore, $T_c(N)$ can be expressed as the summation of numbers of cell pairs to resolve in all levels of the tree starting from DM_a:

$$\begin{aligned}T_c(N) &= I + Is^{2d-1} + Is^{2(2d-1)} + \cdots + Is^{n(2d-1)}\\ &= \frac{I\left[s^{(2d-1)(n+1)}-1\right]}{s^{2d-1}-1}\end{aligned} \tag{15}$$

where n is the total number of levels in the tree visited by the algorithm. The value of n increases by 1 when N increases to s^dN. Therefore, by revisiting Eq. (15), we have the following recurrence:

$$T_c(s^dN)=\frac{I\left[s^{(2d-1)(n+2)}-1\right]}{s^{2d-1}-1}=s^{2d-1}T_c(N)-o(1) \tag{16}$$

Based on the master theorem [10], the above recurrence gives

$$T_c(N)=\Theta\big(N^{\log_{s^d}s^{2d-1}}\big)=\Theta\big(N^{\frac{2d-1}{d}}\big).$$

■

Now let us investigate the time complexity for performing operation (ii), i.e., pairwise distance calculation. We have similar results as in Theorem 1.

Theorem 2: With reasonable spatial distribution of particles in the simulated system, the time spent by DM-SDH on calculating pairwise distances of particles in non-resolvable leaf nodes is $\Theta\big(N^{\frac{2d-1}{d}}\big)$.

Proof: As shown in the derivation of Eq. (16), there are $Is^{n(2d-1)}$pairs of leaf nodes to resolve, among which $Is^{n(2d-1)}\frac{1}{s} = Is^{n(2d-1)-1}$ will be unresolved and the pairwise distances of the particles in them need to be computed one by one. When system size increases from N to s^dN, the number of unresolved leaf node pairs (denoted as L) becomes $Is^{(n+1)(2d-1)-1}$. Thus, we get the following recurrence:

$$L(s^dN)=s^{2d-1}L(N),$$

which is essentially the same as Eq. (16) and we easily get

$$L(N)=\Theta\big(N^{\frac{2d-1}{d}}\big).$$

To prove Theorem 2, we need to transform the above results into the number of distance calculations in the unresolved leaf nodes. This extension is obviously true for uniformly distributed particles, in which the expected number of particles in a cell is proportional to the cell size. However, uniform distribution is not necessary for Theorem 2 to be true.

Let us consider any pair of non-resolvable cells $\mathcal{A}$ (with particle count a) and $\mathcal{B}$ (with particle count b) on the leaf level DM_k of the tree. Note that we cannot say $a = b$ (due to the non-uniform data distribution), and we expect to have $T_k = ab$

distances to calculate for these two cells. When the system size increases from N to s^dN, we build another level of density map DM_{k+1}, in which $\mathcal{A}$ and $\mathcal{B}$ are both divided into s^d cells. Let us denote the original number of particles in the subcells as a_i ($i \in \{1, 2, \cdots, s^d\}$) and b_j ($j \in \{1, 2, \cdots, s^d\}$). We immediately have $a = \sum_{i=1}^{s^d} a_i$ and $b = \sum_{j=1}^{s^d} b_j$. When N increases to s^dN, a_i and b_j all get a s^d-fold increase and the expected number of distance calculations becomes

$$T_{k+1} = \sum_{i,j} P_{i,j} s^d a_i s^d b_j \tag{17}$$

where $P_{i,j}$ is a binary variable that tells whether subcells i and j are non-resolvable on DM_{k+1}. Without any assumptions, we only know that the average of $P_{i,j}$ over all combinations of i and j is $\frac{1}{s}$ (Lemma 2). For Theorem 2 to be true, we need to show that $T_{k+1} = \frac{s^{2d}}{s} T_k = s^{2d-1}ab$.

We first show that, if the distribution of particles is *cell-wise uniform* on density map DM_k, we can achieve the above condition. Being cell-wise uniform means that the data are uniformly distributed within each cell, i.e., we have $a_1 = a_2 = \cdots = a_{s^d} = \frac{a}{s^d}$ and $b_1 = b_2 = \cdots = b_{s^d} = \frac{b}{s^d}$, which easily leads to $T_{k+1} = \frac{1}{s} \sum P_{i,j} s^d a s^d b = s^{2d-1}ab$. Being a weaker assumption than system-wise uniform distribution (which also requires $a = b$), the cell-wise uniform distribution is a safe assumption in particle simulations - particles will not be arbitrarily clustered due to the existence of bonds or inter-particle forces [11], [12]. Note that we only need to make this assumption in the leaf nodes. Cell-wise uniform is also a popular assumption in current spatial-temporal database studies [13]. In Section IV-F of [14], we further show that Theorem 2 holds true under more types of spatial distributions that are reasonable in simulation data. The above reasoning can be easily extended to 3D data and tiling factor $s > 2$. ■

Putting Theorem 1 and Theorem 2 together, we conclude that **the time complexity of DM-SDH is** $\Theta\left(N^{\frac{2d-1}{d}}\right)$. We have mentioned that our analysis is done based on an over-estimation of the coverable regions on each density map, and the estimation error decreases as m increases. Relate this to Lemma 2, we have an underestimated non-covering factor α on each level. Since the estimation is more accurate on larger m, the real ratio of $\alpha(m+1)$ to $\alpha(m)$ can only be smaller than the one given by Lemma 2. This means that the results shown in both theorems become an upper bound. As a result, complexity of the algorithm becomes $O\left(N^{\frac{2d-1}{d}}\right)$.

Note that the time complexity has nothing to do with the tiling factor s. In practice, we prefer smaller s values. Recall that the first map DM_a should be the first level with cell size $\delta \leq p/\sqrt{d}$. With a bushy tree as a result of large s value, the cell size increases more dramatically and we could end up a DM_a with cell size way smaller than $p/\sqrt{d}$, giving rise to more cells to resolve (Eq. (15)).

VI. Conclusions and Future Work

In this paper, we present analytical results related to the time complexity of a Quad tree-based algorithm for computing the spatial distance histogram of large scale spatial dataset. Being the main building blocks of high-level analytics in particle simulations, such histograms are of great importance in domain-specific hypothesis testing and scientific discovery. This paper focuses on the methodology we adopt to accomplish the analysis: we transform the problem into quantifying the area of certain regions in space such that geometric modeling can be used to generate rigorous results. Our analysis shows that the recently proposed algorithm has complexity $O(N^{\frac{3}{2}})$ for 2D data and $O(N^{\frac{5}{3}})$ for 3D data, which beat the quadratic brute-force algorithm. We also show that the conclusion holds true with reasonable assumptions made on the spatial distribution of particles in the simulated system.

Immediate future work in this area involves space partitioning methods with cell shapes other than square (e.g., rectangles, triangles). Our conjecture here is that the choice of shape will not affect the main results presented in this paper (e.g., Lemma 2). Therefore, a generalized model is needed to describe this phenomenon. One paradigm skipped by this paper (due to space limitations) is the approximate algorithm [6] derived from the DM-SDH algorithm. While experimental results show very promising tradeoffs of running time and query error, probabilistic models have to be developed to study tight bounds of the error.

References

[1] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithm to Applications*, ser. Computational Science Series. Academic Press, 2002, vol. 1.

[2] M. P. Allen and D. J. Tildesley, *Computer Simulations of Liquids*. Clarendon Press, Oxford, 1987.

[3] J. L. Stark and F. Murtagh, *Astronomical Image and Data Analysis*. Springer, 2002.

[4] S. Klasky, B. Ludaescher, and M. Parashar, "The Center for Plasma Edge Simulation Workflow Requirements," in *EEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow'06)*, 1991, pp. 73–73.

[5] A. Filipponi, "The radial distribution function probed by X–ray absorption spectroscopy," *J. Phys.: Condens. Matter*, vol. 6, pp. 8415–8427, 1994.

[6] Y.-C. Tu, S. Chen, and S. Pandit, "Computing Distance Histograms Efficiently in Scientific Databases," in *Proceedings of International Conference on Data Engineering (ICDE)*, March 2009.

[7] A. G. Gray and A. W. Moore, "N-body problems in statistical learning," in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2000, pp. 521–527.

[8] J. A. Orenstein, "Multidimensional Tries used for Associative Searching," *Information Processing Letters*, vol. 14, no. 4, pp. 150–157, 1982.

[9] H. Samet, "The quadtree and related hierarchical data structures," *ACM Comput. Surv.*, vol. 16, no. 2, pp. 187–260, 1984.

[10] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. MIT Press and McGraw-Hill, 2001, pp. 73–75.

[11] M. Allen, *Introduction to Molecular Dynamics Simulation*. vol. 23: John von Neumann Institute of Computing, NIC Seris, 2003.

[12] A. Omeltchenko, T. J. Campbell, R. K. Kalia, X. Liu, A. Nakano, and P. Vashishta, "Scalable I/O of Large-Scale Molecular Dynamics Simulations: A Data-Compression Algorithm," *Computer Physics Communications*, vol. 131, pp. 78–85, 2000.

[13] Y. Tao, J. Sun, and D. Papadias, "Analysis of predictive spatio-temporal queries," *ACM Trans. Database Syst.*, vol. 28, no. 4, pp. 295–336, 2003.

[14] Y.-C. Tu, S. Chen, and S. Pandit, "Computing Spatial Distance Histograms Efficiently in Scientific Databases," Department of Computer Science and Engineering, University of South Florida, Tech. Rep. CSE/08-103, http://www.cse.usf.edu/~ytu/pub/tr/pdh.pdf, 2008.

APPENDIX I
THE DERIVATION OF EQ. (13)

We accomplish this proof by studying the difference between $\frac{A(m)}{B(m)}$ and $\frac{1}{2}$. First, we see

$$\begin{aligned} A(m) - \frac{B(m)}{2} &= \sum_{i=2}^{l} \sqrt{2(i-1)^2 - \frac{1}{4}} - 4\sum_{i=2}^{l} \theta_{m+1}\sqrt{2(i-1)^2 - \theta_{m+1}^2} \\ &+ 2\sum_{i=2}^{l} \theta_m \sqrt{2(i-1)^2 - \theta_m^2} + 8\sum_{i=2}^{l} (i-1)^2 \arctan \frac{\sqrt{8(i-1)^2 - \theta_{m+1}^2}}{\theta_{m+1}} \\ &- 4\sum_{i=2}^{l} (i-1)^2 \arctan \frac{\sqrt{8(i-1)^2 - \theta_m^2}}{\theta_m} - 4\sum_{i=2}^{l} (i-1)^2 \arctan \sqrt{8(i-1)^2 - 1} \end{aligned} \tag{18}$$

When $l \to \infty$, we have the results shown in (19).

$$\begin{aligned} \sum_{i=2}^{l} \sqrt{2(i-1)^2 - \frac{1}{4}} &\longrightarrow \sum_{i=2}^{l} \sqrt{2}(i-1) \\ \sum_{i=2}^{l} \theta_{m+1}\sqrt{2(i-1)^2 - \theta_{m+1}^2} &\longrightarrow \sum_{i=2}^{l} \theta_{m+1}\sqrt{2}(i-1) \\ \sum_{i=2}^{l} \theta_m \sqrt{2(i-1)^2 - \theta_m^2} &\longrightarrow \sum_{i=2}^{l} \theta_m \sqrt{2}(i-1) \\ \sum_{i=2}^{l} (i-1)^2 \arctan \frac{\sqrt{8(i-1)^2 - \theta_{m+1}^2}}{\theta_{m+1}} &\longrightarrow \sum_{i=2}^{l} (i-1)^2 \arctan 2\sqrt{2}(i-1) \\ \sum_{i=2}^{l} (i-1)^2 \arctan \frac{\sqrt{8(i-1)^2 - \theta_m^2}}{\theta_m} &\longrightarrow \sum_{i=2}^{l} (i-1)^2 \arctan 2\sqrt{2}(i-1) \\ \sum_{i=2}^{l} (i-1)^2 \arctan \sqrt{8(i-1)^2 - 1} &\longrightarrow \sum_{i=2}^{l} (i-1)^2 \arctan 2\sqrt{2}(i-1) \end{aligned} \tag{19}$$

Plugging the left-hand side of six formulae in (19) into Eq. (18), we get $A(m) - \frac{B(m)}{2} \longrightarrow 0$ and thus $A(m) \longrightarrow \frac{B(m)}{2}$.

APPENDIX II
RELEVANT QUANTITIES IN 3D ANALYSIS

These formulae are listed on the last page of this paper as Eq. (20) to Eq. (22).

$$\begin{aligned} V_{\mathbf{B}'}(m) &= \iint_{\mathbf{B}'} dxdy \int_{\frac{\delta}{2}}^{\sqrt{p^2 - \left(x - \frac{\delta}{2^m}\right)^2 - \left(y - \frac{\delta}{2^m}\right)^2} + \frac{\delta}{2^m}} dz \\ &= \iint_{\mathbf{B}'} \left[\sqrt{p^2 - \left(x - \frac{\delta}{2^m}\right)^2 - \left(y - \frac{\delta}{2^m}\right)^2} - \delta\theta_m \right] dxdy \\ &= \int_a^{\frac{\pi}{4}} d\phi \int_b^c \left(\sqrt{p^2 - r^2} - \delta\theta_m \right) r dr \\ &= \int_a^{\frac{\pi}{4}} \left[-\frac{1}{3}(p^2 - r^2)^{\frac{3}{2}} - \frac{\delta\theta_m}{2} r^2 \right] \Bigg|_b^c d\phi \\ &= \int_a^{\frac{\pi}{4}} \left[-\frac{(\delta\theta_m)^3}{3} + \frac{1}{3}\left(p^2 - b^2\right)^{\frac{3}{2}} - \frac{\delta\theta_m}{2}\left[p^2 - (\delta\theta_m)^2\right] + \frac{\delta\theta_m}{2} b^2 \right] d\phi \end{aligned} \tag{20}$$

Here we have $a = \arctan \frac{\delta\theta_m}{\sqrt{p^2 - 2(\delta\theta_m)^2}}$, $b = \frac{\delta\theta_m}{\sin\phi}$, and $c = \sqrt{p^2 - (\delta\theta_m)^2}$.

The coverable region is

$$f(i,m)=\begin{cases}\frac{4}{3}\pi p^3+6p\left(\delta-\frac{2\delta}{2^m}\right)^2+3\pi p^2\delta\left(\delta-\frac{2\delta}{2^m}\right)+\left(\delta-\frac{2\delta}{2^m}\right)^3 & n=1,m\geq 1\\ \frac{4}{3}\pi(ip)^3-\frac{4}{3}\pi[(i-1)p]^3 & n\geq 2,m=1\\ \frac{4}{3}\pi(ip)^3+6ip\left(\delta-\frac{2\delta}{2^m}\right)^2+3\pi(ip)^2\delta\left(\delta-\frac{2\delta}{2^m}\right)+\left(\delta-\frac{2\delta}{2^m}\right)^3-v(i,m,p,\delta) & n\geq 2,m>1\end{cases}$$

Simplifying the above with $p=\sqrt{3}\delta$, we get

$$f(i,m)=\begin{cases}\left[4\sqrt{3}\pi+6\sqrt{3}\left(1-\frac{2\delta}{2^m}\right)^2+9\pi\left(1-\frac{2}{2^m}\right)+\left(1-\frac{2}{2^m}\right)^3\right]\delta^3 & n=1,m\geq 1\\ [4\sqrt{3}\pi i^3-4\sqrt{3}\pi(i-1)^3]\delta^3 & n\geq 2,m=1\\ \left[4\sqrt{3}\pi i^3+6\sqrt{3}i\left(1-\frac{2\delta}{2^m}\right)^2+9\pi i^2\left(1-\frac{2}{2^m}\right)+\left(1-\frac{2}{2^m}\right)^3-v(i,m,p)\right]\delta^3 & n\geq 2,m>1\end{cases}$$

where $v(i,m,p,\delta)=16V_{\mathbf{B}'}(m)$ and $v(i,m,p,\delta)=\delta^3 v(i,m,p)$.
Continue with the same reasoning as in Section III-C, we have

$$\begin{aligned}G(l) &= \frac{\sum_{i=1}^{l}g(i)}{\delta^3}=\sum_{i=1}^{l}\left(4\sqrt{3}\pi i^3+6\sqrt{3}i+9\pi i^2+1\right)\\ &\quad-16\sum_{i=2}^{l}\int_q^{\frac{\pi}{4}}\left[-\frac{1}{24}+\frac{1}{3}\left(3(i-1)^2-\left(\frac{1}{2\sin\phi}\right)^2\right)^{\frac{3}{2}}-\frac{1}{4}\left(3(i-1)^2-\left(\frac{1}{2}\right)^2\right)+\frac{1}{16}\frac{1}{(\sin\phi)^2}\right]d\phi \quad (21)\end{aligned}$$

where $q=\arctan\frac{\frac{1}{2}}{\sqrt{3(i-1)^2-2\left(\frac{1}{2}\right)^2}}$, and the following formulae for the accumulated volume for all coverable regions F.

$$\frac{\sum_{i=1}^{l}f(i,m)}{\delta^3}=\begin{cases}4\sqrt{3}\pi+\sum_{i=2}^{l}\left[4\sqrt{3}i^3-4\sqrt{3}\pi(i-1)^3\right]=4\sqrt{3}\pi l^3\ , & m=1\\ \sum_{i=1}^{l}\left[4\sqrt{3}\pi i^3+6\sqrt{3}i\left(1-\frac{2}{2^m}\right)^2+9\pi i^2\left(1-\frac{2}{2^m}\right)+\left(1-\frac{2}{2^m}\right)^3\right]\\ \quad -16\sum_{i=2}^{l}\int_s^{\frac{\pi}{4}}\left[-\frac{\theta_m^3}{3}+\frac{1}{3}\left(3-\left(\frac{\theta_m}{\sin\phi}\right)^2\right)^{\frac{3}{2}}-\frac{\theta_m}{2}\left(3-\theta_m^2\right)+\frac{\theta_m}{2}\left(\frac{\theta_m}{\sin\phi}\right)^2\right]d\phi\ , & m>1\end{cases} \quad (22)$$

in which $s=\arctan\frac{\theta_m}{\sqrt{3(i-1)^2-2\theta_m^2}}$.

APPENDIX III
PROOF OF LEMMA 2

Proof: Proof is accomplished in a similar way to that of Lemma 1. While the total area of all bucket regions Eq. (8) is still the same, Eq. (9) and Eq. (10) become the following equation for all $m \geq 1$:

$$\begin{aligned} F(l,m,s) &= \frac{\sum_{i=1}^{l} f(i,m,s)}{\delta^2} \\ &= \sum_{i=1}^{l}\left[\pi(ip)^2 + 4ip\left(\delta - \frac{2\delta}{s^m}\right) + \left(\delta - \frac{2\delta}{s^m}\right)^2\right] \\ &\quad -\sum_{i=2}^{l}(i-1)^2\left[8\arctan\frac{\sqrt{2(i-1)^2 - {\theta'_m}^2}}{{\theta'_m}^2} - 2\pi\right] + 4\sum_{i=2}^{l}\left[\theta'_m\sqrt{2(i-1)^2 - {\theta'_m}^2} - {\theta'_m}^2\right], \end{aligned} \tag{23}$$

which gives $\frac{\alpha(m+1,s)}{\alpha(m,s)} = \frac{A(m,s)}{B(m,s)}$ where

$$\begin{aligned} A(m,s) &= 1 + \frac{4\sqrt{2}(l+l^2)}{s^{1+m}} - l\left(1 - \frac{2}{s^{1+m}}\right)^2 + 4(l-1)\left(\frac{1}{2} - \frac{1}{s^{1+m}}\right)^2 \\ &\quad -4\sum_{i=2}^{l}\theta'_{m+1}\sqrt{2(i-1)^2 - {\theta'_{m+1}}^2} + 8\sum_{i=2}^{l}(i-1)^2\arctan\frac{\sqrt{2(i-1)^2 - {\theta'_{m+1}}^2}}{\theta'_{m+1}} \\ &\quad +\sum_{i=2}^{l}\sqrt{8(i-1)^2-1} - 8\sum_{i=2}^{l}(i-1)^2\arctan\sqrt{8(i-1)^2-1} \end{aligned} \tag{24}$$

and

$$\begin{aligned} B(m,s) &= 1 + \frac{4\sqrt{2}(l+l^2)}{s^{m}} - l\left(1 - \frac{2}{s^{m}}\right)^2 + 4(l-1)\left(\frac{1}{2} - \frac{1}{s^{m}}\right)^2 \\ &\quad -4\sum_{i=2}^{l}\theta'_{m}\sqrt{2(i-1)^2 - {\theta'_{m}}^2} + 8\sum_{i=2}^{l}(i-1)^2\arctan\frac{\sqrt{2(i-1)^2 - {\theta'_{m}}^2}}{\theta'_{m}} \\ &\quad +\sum_{i=2}^{l}\sqrt{8(i-1)^2-1} - 8\sum_{i=2}^{l}(i-1)^2\arctan\sqrt{8(i-1)^2-1} \end{aligned} \tag{25}$$

Following the reasoning in Appendix I, we compare the value of $\frac{A(m,s)}{B(m,s)}$ to $\frac{1}{s}$. And we have

$$\begin{aligned} A(m,s)s - B(m,s) &= (s-1)\sum_{i=2}^{l}\sqrt{8(i-1)^2-1} - 8(1-s)\sum_{i=2}^{l}(i-1)^2\arctan\sqrt{8(i-1)^2-1} \\ &\quad -4(1-s)\sum_{i=2}^{l}\theta'_{m+1}\sqrt{2(i-1)^2 - {\theta'_{m+1}}^2} \\ &\quad +8(s-1)\sum_{i=2}^{l}(i-1)^2\arctan\frac{\sqrt{2(i-1)^2 - {\theta'_{m+1}}^2}}{\theta'_{m+1}} \end{aligned} \tag{26}$$

When $l \to \infty$, we have the results shown in (27).

$$\begin{aligned} \sum_{i=2}^{l}\sqrt{2(i-1)^2 - {\theta'_{m+1}}^2} &\longrightarrow \frac{1}{2}\sum_{i=2}^{l}\sqrt{8(i-1)^2-1} \\ \sum_{i=2}^{l}(i-1)^2\arctan\frac{\sqrt{8(i-1)^2 - {\theta'_{m+1}}^2}}{{\theta'}_{m+1}} &\longrightarrow \sum_{i=2}^{l}(i-1)^2\arctan\sqrt{8(i-1)^2-1} \end{aligned} \tag{27}$$

Plugging the left-hand side of the above two formulae in (27) into Eq. (26), we get $sA(m,s) - B(m,s) \longrightarrow 0$ and thus $A(m,s) \longrightarrow \frac{B(m,s)}{s}$. ■