Complexity of estimating multi-way join result sizes for area skewed spatial data

Ho-Hyun Park, Chin-Wan Chung

Department of Computer Science, Korea Advanced Institute of Science and Technology, 373-1, Kusong-dong, Yusong-gu, Taejon 305-701, South Korea

Received 25 March 2000; received in revised form 10 July 2000
Communicated by K. Iwama

Abstract

In a real life environment, spatial data is highly skewed. In general, there are two kinds of skews in spatial data. One is the placement skew and the other is the area skew. This paper introduces methods and the complexity of estimating the result sizes of the multi-way join for the area skewed spatial data. Especially, this paper describes the number and sort of the statistics which the optimizer should keep in order to calculate the multi-way join result size. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Spatial databases; Multi-way spatial join; Spatial join selectivity; Area skew; Databases

1. Introduction

The spatial join is a common spatial query type which requires high processing cost due to the high complexity and large volume of spatial data. Therefore, the exact estimation of the result size of the spatial join has a great influence on the query optimizer and the spatial database management system. The \( n \)-way \( (n \geq 2) \) spatial join combines \( n \) spatial relations using \( n - 1 \) or more spatial predicates. An example of a 3-way spatial join is “Find all buildings which are adjacent to roads that intersect with boundaries of districts.”

In a real life environment, spatial data is highly skewed. In general, there are two kinds of skews in spatial data. One is the placement skew and the other is the area skew [1]. Recently, under the assumption of the uniform placement distribution, the formula of estimating the result size of the multi-way spatial join for two kinds of query types, tree and clique, were developed [5]. However, the formula is applied only in some special cases of the data areas. This paper extends the formula to the arbitrary area of spatial data, and then analyzes the characteristics and complexities of the extended formula. Especially, we describe the number and sort of the statistics which the optimizer should keep in order to calculate the join result size from the formula for each join query type.
Table 1
Notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r_i$</td>
<td>relation $i$</td>
</tr>
<tr>
<td>$s_{ri,j}$</td>
<td>$j$th object in $r_i$</td>
</tr>
<tr>
<td>$s_{ri,j,x}$</td>
<td>$x$-length of $j$th object in $r_i$</td>
</tr>
<tr>
<td>$s_{ri,j,y}$</td>
<td>$y$-length of $j$th object in $r_i$</td>
</tr>
<tr>
<td>$s_{ri,j,xkyl}$</td>
<td>$(x$-length)$^k * (y$-length)$^l$ of $j$th object in $r_i$</td>
</tr>
<tr>
<td>$D_{sk}ri;j;x$</td>
<td>$x$-length of query window $q$</td>
</tr>
<tr>
<td>$D_{sk}ri;j;y$</td>
<td>$y$-length of query window $q$</td>
</tr>
<tr>
<td>$N_{ri}$</td>
<td>number of objects in $r_i$</td>
</tr>
<tr>
<td>$L_{ri,x}$</td>
<td>average $x$-length for all objects in $r_i$</td>
</tr>
<tr>
<td>$L_{ri,y}$</td>
<td>average $y$-length for all objects in $r_i$</td>
</tr>
</tbody>
</table>

2. Background

In this paper, we consider spatial objects only in the two-dimensional unit space $[0, 1]^2$, and assume that the objects are rectangular, i.e., consider only the MBR (Minimum Bounding Rectangle) of a real object. The notations to be used in this paper are summarized in Table 1.

Under the assumption that the placement distribution is uniform, the formula for estimating the result size of the window query on $r_i$ is shown in the following formula [3,4]:

$$ Size(\sigma_q(r_i)) = \sum_{j=1}^{N_{ri}} (s_{ri,j,x} + q_x) * (s_{ri,j,y} + q_y). $$  \hspace{1cm} (1)

In the above formula, a query optimizer does not know object information such as $s_{ri,j,x}$ and $s_{ri,j,y}$ in optimization time. For the estimation of the query result size, a query optimizer generally keeps statistics such as $N_{ri}$, $L_{ri,x}$, $L_{ri,y}$ and $L_{ri,xy}$ in the system catalog. Therefore, we must transform the above formula into the expression in terms of statistics as follows [3]:

$$ Size(\sigma_q(r_i)) = L_{ri,x} + q_x * L_{ri,y} + q_y * N_{ri} * q_x * q_y. $$  \hspace{1cm} (2)

The following shows the formula for estimating the result size of the spatial join between $r_i$ and $r_j$ under the uniform assumption of the placement distribution [2,6]:

$$ Size(r_i \bowtie_{\text{intersect}} r_j) = \sum_{k=1}^{N_{ri}} \sum_{l=1}^{N_{rj}} (s_{ri,k,x} + s_{rj,l,x}) * (s_{ri,k,y} + s_{rj,l,y}). $$  \hspace{1cm} (3)

The following expression is the transformation of the above formula in terms of the expression of statistics:

$$ Size(r_i \bowtie_{\text{intersect}} r_j) = L_{ri,x} * N_{rj} + L_{ri,y} * N_{ri} + L_{rj,x} + L_{rj,y} * L_{ri,x} + N_{ri} * L_{rj,xy}. $$  \hspace{1cm} (4)

---

2 This paper can be applied in both the relational model and the object-oriented model. However, we use the terminology relation for notational familiarity instead of class or extent.
Eqs. (2) and (4) do not assume anything about the area of spatial data, i.e., the above expressions can be applied to the arbitrary area distribution of spatial data. In Eqs. (2) and (4), we need four statistics per relation to estimate the result size of the window query and the spatial join (e.g., for \( r_i, N_{r_i}, L_{r_i,x}, L_{r_i,y} \) and \( L_{r_i,xy} \)). If we assume that the \( x \)-length distribution is independent of the \( y \)-length distribution, we need only three statistics per relation because \( L_{r_i,xy} \) can be represented by \( L_{r_i,x} \cdot L_{r_i,y} / N_{r_i} \). (Remind that if random variables \( X \) and \( Y \) are mutually independent, \( E(XY) = E(X) \cdot E(Y) \).)

3. Result sizes for multi-way spatial joins

First, we define some terminologies. A term is a component of an expression. An atomic term is an atomic component of an expression such as \( s_{r_i,j,x}, q_x \) and \( L_{r_i,xy} \) in Eqs. (1) and (2). Among atomic terms, object information \( s_{r_i,j,x,y} \) is called the object term, and statistics information \( L_{r_i,x,y} \) the statistics term. A compound term is a composition of terms by operators such as + and *. For example, \( s_{r_i,j,x+y} \) and \( q_x \cdot s_{r_i,j,y} \) are compound terms. Among compound terms, if the outer operators are +, the term is called the addition term, and if *, the production term. For example, Eq. (4) is an addition term which consists of four production terms each of which consists of two statistics terms.

The multi-way spatial join can be modeled by a query graph whose node represents a relation and edge represents a spatial relationship. We treat only the tree type (acyclic query graph) and the clique type (complete query graph) as components of an expression such as \( s_{r_i,j,x,y} \). For example, Eq. (4) is an addition term which consists of four production terms each of which consists of two statistics terms.

In Eqs. (2) and (4), the above formula is an extension of the formula of the 2-way join. However, this extension has a serious restriction. Eq. (5) first approximates the \( x \)- and \( y \)-lengths of all data to the average value per axis. As it will be explained later in this section, the above formula is satisfied only in the case that the \( x \)- and \( y \)-lengths of all data per relation are equal. If we extend Eq. (3) without an approximation, we obtain the following formula:

\[
\text{Size}(Q) = \prod_{k=1}^{n} N_{r_k} \prod_{v_{i,j} : Q(r_i,r_j) = 1} (\tilde{s}_{r_i,x} + \tilde{s}_{r_j,x}) \cdot (\tilde{s}_{r_i,y} + \tilde{s}_{r_j,y}).
\]  

(5)

In the above formula, \( Q \) is a query graph which represents a tree typed multi-way spatial join among relations \( r_1, \ldots, r_n \), and \( Q(r_i,r_j) = 1 \) means there is an intersection relationship between \( r_i \) and \( r_j \). According to Papadias et al. [5], the above formula is an extension of the formula of the 2-way join. However, this extension has a serious restriction. Eq. (5) first approximates the \( x \)- and \( y \)-lengths of all data to the average value per axis. As it will be explained later in this section, the above formula is satisfied only in the case that the \( x \)- and \( y \)-lengths of all data per relation are equal. If we extend Eq. (3) without an approximation, we obtain the following formula:

\[
\text{Size}(Q) = \prod_{k_1=1}^{N_{r_1}} \cdots \prod_{k_n=1}^{N_{r_n}} \prod_{v_{i,j} : Q(r_i,r_j) = 1} (s_{r_i,k_i,x} + s_{r_j,k_j,x}) \cdot (s_{r_i,k_i,y} + s_{r_j,k_j,y}).
\]  

(6)

In the above formula, the inner product term \( \prod_{v_{i,j} : Q(r_i,r_j) = 1} (s_{r_i,k_i,x} + s_{r_j,k_j,x}) \cdot (s_{r_i,k_i,y} + s_{r_j,k_j,y}) \) is the probability with which an arbitrary \( n \)-tuple \( s_{r_1,k_1}, \ldots, s_{r_n,k_n} \) from relations \( r_1, \ldots, r_n \) satisfies the query \( Q \). We call the probability the selectivity \( S \) for \( Q \). For the estimation of the query result size, Eq. (6) should also be derived using the statistics terms. If we derive the formula of Eq. (6) using the statistics terms in a 3-way join \( (r_1 \bowtie r_2 \text{ and } r_1 \bowtie r_3) \), the following expression is obtained:

\[
\text{Size}(Q_3) = L_{r_1,x}^2 \cdot y^2 \cdot N_{r_1} \cdot L_{r_1,y}^2 \cdot N_{r_2} \cdot L_{r_2,y} \cdot N_{r_3} \cdot L_{r_3,y} \cdot L_{r_1,x} \cdot L_{r_2,y} \cdot L_{r_3,y} + L_{r_1,x}^2 \cdot y^2 \cdot N_{r_3} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_2,x} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_2,x}^2 \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_2,x} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_1,x} \cdot L_{r_2,x} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_1,x} \cdot L_{r_2,x} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_1,x} \cdot L_{r_2,x} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y} + L_{r_1,x} \cdot L_{r_2,x} \cdot L_{r_2,y} \cdot L_{r_3,x} \cdot L_{r_3,y}.
\]  

(7)
Eq. (7) is more complex than Eqs. (2) and (4). The statistics for \( r_2 \) and \( r_3 \) are just \( N_{r_1}, L_{r_1} \times, L_{r_1} \times, L_{r_2} \times \times, L_{r_3} \times \times, L_{r_4} \times \times, L_{r_5} \times \times, L_{r_6} \times \times \times, L_{r_7} \times \times \times, L_{r_8} \times \times \times \times \times, L_{r_9} \times \times \times \times \times \times, L_{r_{10}} \times \times \times \times \times \times \times, L_{r_{11}} \times \times \times \times \times \times \times \times, \) and the second question is concerned with the permanent space complexity to keep the statistics information. The first question is to derive the formula of Eq. (6) using statistics terms, how many terms do we have? And how many statistics should we keep per relation in order to estimate the result size of the tree typed multi-way spatial join? The addition term which consists of 4-way joins into the statistics terms, the resulting expression is an addition term which consists of 4 terms in total:

\[
\sum \frac{d}{d} \left( L_{r_1} \times \times, L_{r_2} \times \times, L_{r_3} \times \times, L_{r_4} \times \times, L_{r_5} \times \times, L_{r_6} \times \times, L_{r_7} \times \times \times, L_{r_8} \times \times \times, L_{r_9} \times \times \times \times, L_{r_{10}} \times \times \times \times, L_{r_{11}} \times \times \times \times \times \times, L_{r_{12}} \times \times \times \times \times \times \times, L_{r_{13}} \times \times \times \times \times \times \times \times, \right)
\]

Eq. (8) has much more terms than Eq. (7) and more statistics for \( r_1 \). As we show in Eqs. (7) and (8), if we transform the formula of object terms for a tree typed multi-way spatial join into that of statistics terms, the resulting expression is a complex addition term of production terms of statistics terms. Now, we have two questions. If we derive the formula of Eq. (6) using statistics terms, how many terms do we have? And how many statistics should we keep per relation in order to estimate the result size of the tree typed multi-way spatial join? The first question is concerned with the time and the temporary space complexity for the evaluation of Eq. (6) using statistics terms, and the second question is concerned with the permanent space complexity to keep the statistics information. The following two lemmas give the answers.

**Lemma 1.** If we transform Eq. (6) into the statistics terms, the resulting expression is an addition term which consists of \( 4^{n-1} \) distinct production terms each of which consists of distinct \( n \) statistics terms each of which is from \( r_i \) (\( 1 \leq i \leq n \)).

**Proof.** We first show that if we derive the selectivity \( S \) for \( Q \) using object terms, the resulting expression is an addition term which consists of \( 4^{n-1} \) distinct production terms each of which consists of distinct \( n \) object terms each of which is from \( r_i \) (\( 1 \leq i \leq n \)).

(i) For the case of \( n = 2 \). Let two relations to be joined be \( r_i \) and \( r_j \). For any objects \( s_{r_i,k_i} \) and \( s_{r_j,k_j} \), the probability \( (S_2) \) that the two objects mutually intersect is represented by the following expression:

\[
S_2 = (s_{r_i,k_i}, x + s_{r_j,k_j}, x) \ast (s_{r_i,k_i}, y + s_{r_j,k_j}, y)
\]

\[
= s_{r_i,k_i}, x \ast s_{r_i,k_i}, y + s_{r_i,k_i}, x \ast s_{r_j,k_j}, y + s_{r_j,k_i}, x \ast s_{r_j,k_j}, y
\]

\[
= s_{r_i,k_i}, x \ast y + s_{r_j,k_i}, x \ast y + s_{r_i,k_j}, y \ast s_{r_j,k_j}, x + 1 \ast s_{r_j,k_j}, x.
\]

(9)
In Eq. (9), since the two 1’s can be represented by \( s_{r_1,k_1,x^0y^0} \) and \( s_{r_1,k_1,x^0y^0} \), respectively, they can be considered as object terms. Obviously, Eq. (9) is an addition term which consists of 4 distinct production terms each of which consists of 2 object terms from \( r_i \) and \( r_j \), respectively.

(ii) For the case of \( n = m > 2 \): Let \( Q_m \) be a subquery of \( Q \) consisting of \( r_1 \) to \( r_m \). Let \( S_m \) be the probability that any objects \( s_{r_1,k_1}, \ldots, s_{r_m,k_m} \) satisfy \( Q_m \) and be an addition term which consists of \( 4^{m-1} \) distinct production terms each of which consists of \( m \) distinct object terms each of which is from \( r_i \) (\( 1 \leq i \leq m \)). Let the next relation to be joined with a relation in \( Q_m \) be \( r_{m+1} \), and \( r_i \) (\( 1 \leq i \leq m \)) be the relation to be joined with \( r_{m+1} \). Then, the selectivity for the \((m+1)\)-way join is the following:

\[
S_{m+1} = S_m \ast (s_{r_i,k_i,x} + s_{r_{m+1},k_{m+1},x}) \ast (s_{r_i,k_i,y} + s_{r_{m+1},k_{m+1},y}) \\
= S_m \ast (s_{r_i,k_i,xy} + s_{r_i,k_i,k_{m+1},y} + s_{r_i,k_i,y} \ast s_{r_{m+1},k_{m+1},x} + 1 \ast s_{r_{m+1},k_{m+1},xy}) . \tag{10}
\]

Now, \( S_m \) consists of \( 4^{m-1} \) distinct production terms for \( r_1, \ldots, r_m \). And the expression between \( r_i \) and \( r_{m+1} \) consists of 4 distinct production terms for \( r_{m+1} \) which are \( 1, s_{r_m,k_{m+1},y}, s_{r_m,k_{m+1},x} \) and \( s_{r_{m+1},k_{m+1},xy} \). Therefore, \( S_{m+1} \) consists of \( 4^{m-1} \ast 4 = 4^m \) distinct production terms for \( r_1, \ldots, r_{m+1} \). Since all 1’s in Eq. (10) can be represented by the relevant object terms (e.g., for \( r_i, s_{r_i,k_i,x^0y^0} \)), each production term consists of the object terms of \( r_1 \) to \( r_{m+1} \). In this way, if we expand the expression \( S \) for the whole query graph \( Q \), we can get an addition term which consists of \( 4^{n-1} \) distinct production terms each of which consists of \( n \) distinct object terms each of which is from \( r_i \) (\( 1 \leq i \leq n \)).

Since \( \text{Size}(Q) = \sum_{k_1=1}^{N_1} \cdots \sum_{k_m=1}^{N_m} S \), if we apply the summations \( \sum_{k_1=1}^{N_1} \cdots \sum_{k_m=1}^{N_m} \) to \( S \), the only change is that the object terms in \( S \) are converted to the relevant statistics terms in \( \text{Size}(Q) \). Therefore, \( \text{Size}(Q) \) is an addition term which consists of \( 4^{n-1} \) distinct production terms and each production term in \( \text{Size}(Q) \) consists of \( n \) statistics terms for \( r_1 \) to \( r_n \).

In the 5-way join, the statistics expression consists of \( 4^5 = 256 \) production terms, and each production term consists of 5 statistics terms. Therefore, the optimizer needs memory for saving \( 4^5 \ast 5 = 1280 \) statistics terms. Each term must have \( x \)-value and \( y \)-value, and if each value occupies 1 byte, the optimizer needs 2560 bytes. And it also needs computation time for \( 4^4 - 1 = 255 \) additions and \( 4^4 \ast 4 = 1024 \) productions. In the 10-way join, the optimizer needs \( 4^9 \ast 10 = 2,621,440 \) statistics terms, therefore main memory of about 5 MB. And it also needs computation time for \( 4^9 - 1 = 262,143 \) additions and \( 4^9 \ast 9 = 2,359,296 \) productions. In the 15-way join, the optimizer needs \( 4^{14} \ast 15 \) (about 4 billions) statistics terms, therefore main memory of about 8 GB. And it also needs computation time for \( 4^{14} - 1 \) (about 268 millions) additions and \( 4^{14} \ast 14 \) (about 3.8 billions) productions. This incurs tremendous computation time and memory overhead although the memory is for temporary use.

**Lemma 2.** If \( Q \) is a query graph representing a tree typed multi-way spatial join of \( n \) relations and \( r_i \) is a node in \( Q \) representing a relation, and \( \delta_{r_i} \) is the number of edges incident on \( r_i \), the following \((\delta_{r_i} + 1)^2 \) statistics for \( r_i \) (\( 1 \leq i \leq n \)) are needed to estimate the result size of \( Q \):

\[
\begin{array}{cccc}
N_{r_i} & L_{r_i,x} & L_{r_i,x^2} & \cdots & L_{r_i,x^{\delta_{r_i}}} \\
L_{r_i,y} & L_{r_i,xy} & L_{r_i,x^2y} & \cdots & L_{r_i,x^{\delta_{r_i}}y} \\
L_{r_i,y^2} & L_{r_i,xy^2} & L_{r_i,x^2y^2} & \cdots & L_{r_i,x^{\delta_{r_i}}y^2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
L_{r_i,y^{\delta_{r_i}}} & L_{r_i,xy^{\delta_{r_i}}} & L_{r_i,x^2y^{\delta_{r_i}}} & \cdots & L_{r_i,x^{\delta_{r_i}}y^{\delta_{r_i}}}
\end{array}
\]
Proof. As in the case of Lemma 1, we first deal with the selectivity which is represented by object terms.

(i) For the case of $n = 2$: Let two relations to be joined be $r_i$ and $r_j$. The probability ($S_2$) that any two objects $s_{ri,k_i}$ and $s_{rj,k_j}$ from $r_i$ and $r_j$ mutually intersect is the same as Eq. (9). And the number of edges of each relation in the 2-way join is 1. To calculate Eq. (9), $s_{ri,k_i;x}$, $s_{ri,k_i;x^2}$, and 1 for $r_i$ are needed.

(ii) For the case of $n > 2$: Let $Q_m$ be a subgraph of $Q$ consisting of $r_1$ to $r_m$. Let $S_m$ be the probability that any $s_{r_1,k_1}, \ldots, s_{rm,k_m}$ satisfy $Q_m$. Let $d$ be the number of edges of $r_i$ in $Q_m$. Assume that the following object terms for $r_i$ are needed to get $S_m$:

\[
\begin{array}{cccc}
1 & s_{r_i,k_i,x} & s_{r_i,k_i,x^2} & \cdots & s_{r_i,k_i,d} \\
& s_{r_i,k_i,y} & s_{r_i,k_i,xy} & \cdots & s_{r_i,k_i,dy} \\
& s_{r_i,k_i,y^2} & s_{r_i,k_i,xy^2} & \cdots & s_{r_i,k_i,dxy} \\
& \vdots & \vdots & \ddots & \vdots \\
& s_{r_i,k_i,yd} & s_{r_i,k_i,xyd} & \cdots & s_{r_i,k_i,dyd} \\
\end{array}
\]

Let the next relation to be joined with a relation in $Q_m$ be $r_{m+1}$. If the relation in $Q_m$ to be joined with $r_{m+1}$ is $r_i$, $S_{m+1}$ is the same as Eq. (10). In Eq. (10), each production term in the expression between $r_i$ and $r_{m+1}$ has an object term $s_{r_i,k_i,xy}, s_{r_i,k_i,x}, s_{r_i,k_i,y}$ or 1 for $r_i$, and $S_m$ has the above matrix as object terms for $r_i$. Therefore, the object terms for $r_i$ in $S_{m+1}$ are the following:

\[
\begin{array}{cccc}
1 & s_{r_i,k_i,x} & s_{r_i,k_i,x^2} & \cdots & s_{r_i,k_i,d} \\
& s_{r_i,k_i,y} & s_{r_i,k_i,xy} & \cdots & s_{r_i,k_i,dy} \\
& s_{r_i,k_i,y^2} & s_{r_i,k_i,xy^2} & \cdots & s_{r_i,k_i,dxy} \\
& \vdots & \vdots & \ddots & \vdots \\
& s_{r_i,k_i,yd} & s_{r_i,k_i,xyd} & \cdots & s_{r_i,k_i,dyd} \\
& s_{r_i,k_i,yd+1} & s_{r_i,k_i,xyd+1} & \cdots & s_{r_i,k_i,dyd+1} \\
\end{array}
\]

If the relation in $Q_m$ to be joined with $r_{m+1}$ is $r_j$ ($1 \leq i \neq j \leq m$), the expression between $r_j$ and $r_{m+1}$ has nothing to do with $r_i$. Therefore, the object terms for $r_i$ in $S_{m+1}$ are the same as those in $S_m$. Instead, the object terms for $r_j$ increase.

In this way, if we evaluate all relations, we can get $S$ which consists of only object terms. Then, the number of distinct object terms for $r_i$ in $S$ becomes $(d_{r_i} + 1)^2$, i.e., in the above matrix, $d + 1$ becomes $d_{r_i}$. Finally, if we apply the summations ($\sum_{k_i=1}^{N_{r_1}} \cdots \sum_{k_m=1}^{N_{r_m}}$) to $S$ to estimate the result size of $Q$, the object terms for $r_i$ ($1 \leq i \leq n$) are converted to the relevant statistics terms.

In Eq. (7), since the degree of $r_1$ is 2, the statistics for $r_1$ are the following:

\[
\begin{align*}
N_{r_1} & \quad L_{r_1,x} & \quad L_{r_1,x^2} \\
L_{r_1,y} & \quad L_{r_1,xy} & \quad L_{r_1,x^2y} \\
L_{r_1,y^2} & \quad L_{r_1,xy^2} & \quad L_{r_1,x^2y^2} \\
\end{align*}
\]
In Eq. (8), since the degree of $r_1$ is 3, the following $4^2$ statistics are needed:

\[
\begin{align*}
N_{r_1} & \quad L_{r_1,x} \quad L_{r_1,x^2} \quad L_{r_1,x^3} \\
L_{r_1,y} & \quad L_{r_1,xy} \quad L_{r_1,x^2y} \quad L_{r_1,x^3y} \\
L_{r_1,y^2} & \quad L_{r_1,xy^2} \quad L_{r_1,x^2y^2} \quad L_{r_1,x^3y^2} \\
L_{r_1,y^3} & \quad L_{r_1,xy^3} \quad L_{r_1,x^2y^3} \quad L_{r_1,x^3y^3}
\end{align*}
\]

Among the tree typed query graphs with $n$ nodes, if the topology of the query graph is a chain, we need $3^2 = 9$ statistics for $n - 2$ central nodes and $2^2 = 4$ statistics for 2 terminal nodes to estimate the $n$-way join result size. For star topology, we need $n^2$ statistics for 1 central node and $2^2 = 4$ for $n - 1$ terminal nodes. Therefore, in the worst case, $n^2$ statistics are needed for the tree typed query. If there is placement skew in spatial data, the histogram method is popular for the query result size estimation [1,6]. In the method, the uniform placement distribution is assumed per histogram bucket. In such a case, the above statistics are needed per histogram bucket. Since the statistical information is generally cached to main memory for fast selectivity estimation, this incurs much memory overhead. And since we do not know in advance how many relations will be joined, we cannot decide the limit of the number of statistics.

If the $x$-length distribution is independent of the $y$-length distribution, we need only $(2 \cdot \delta_{r_1} + 1)$ statistics such as $N_{r_1}, L_{r_1,x}, L_{r_1,x^2}, \ldots, L_{r_1,x^{\delta_{r_1}}}, L_{r_1,y}, L_{r_1,y^2}, \ldots, L_{r_1,y^{\delta_{r_1}}}$ because $L_{r_1,x^k}y^j = L_{r_1,x^k} \cdot L_{r_1,y^j}/N_{r_1}$. Furthermore, if both $x$- and $y$-length comply known distributions, respectively, we need only 3 statistics such as $N_{r_1}, L_{r_1,x}, L_{r_1,y}$ per relation because the others can be derived using the distribution function and the 3 statistics. However, the $x$- and $y$-length of most real spatial data are not independent and do not comply a specific distribution. If $(\tilde{s}_{r_1,x})^k = L_{r_1,x^k}/N_{r_1}$ and $(\tilde{s}_{r_1,y})^j = L_{r_1,y^j}/N_{r_1}$, Eqs. (5) and (6) are the same. However, this is only the case in which all data lengths for $x$- and $y$-axis are fixed to $\tilde{s}_{r_1,x}$ and $\tilde{s}_{r_1,y}$, respectively.\(^3\)

Now, we consider the result size for the clique typed query. Under the uniform assumption of the placement distribution, Papadias et al. [5] also estimated the result size of the clique typed spatial join as the following formula:

\[
\text{Size}(Q) = \prod_{k=1}^{n} N_{r_k} \ast \left( \sum_{i=1}^{n} \prod_{j=1}^{n} \tilde{s}_{r_j,x} \right) \ast \left( \sum_{i=1}^{n} \prod_{j=1}^{n} \tilde{s}_{r_j,y} \right).
\]

(11)

In the above formula, $Q$ represents a clique typed multi-way spatial join among relations $r_1, \ldots, r_n$. As in the case of Eq. (5), the average data length per axis is first applied. If we also rewrite the above formula without an approximation, we obtain the following formula:

\[
\text{Size}(Q) = \sum_{k_1=1}^{N_{r_1}} \cdots \sum_{k_n=1}^{N_{r_n}} \left( \sum_{i=1}^{n} \prod_{j=1}^{n} s_{r_j,k_{j,x}} \right) \ast \left( \sum_{i=1}^{n} \prod_{j=1}^{n} s_{r_j,k_{j,y}} \right).
\]

(12)

Lemma 3. If we transform Eq. (12) into the statistics terms, the resulting expression is an addition term which consists of distinct $n^2$ production terms and each production term consists of $n$ statistics terms.

Proof. The selectivity $S$ in Eq. (12) is a multiplication of the two addition terms, i.e.,

\[
\left( \sum_{i=1}^{n} \prod_{j=1}^{n} s_{r_j,k_{j,x}} \right) \ast \left( \sum_{i=1}^{n} \prod_{j=1}^{n} s_{r_j,k_{j,y}} \right).
\]

\(^3\) To our knowledge, Papadias et al. [5] made experiments only with the fixed length data.
and each addition term is an expression per axis and has \( n \) distinct production terms. Therefore, if we derive \( S \), \( n^2 \) production terms are generated. For each production term, there is an object term \( s_{r_i,k_i,x^k y^l} (0 \leq k \leq 1, 0 \leq l \leq 1) \) for relation \( r_i \) (of course, 1 is considered to be \( s_{r_i,k_i,x^k y^l} \)). If we apply the summations to \( S \), the object term is converted to the relevant statistics term \( L_{r_i,x^k y^l} \) in \( \text{Size}(Q) \). Therefore each production term in \( \text{Size}(Q) \) consists of \( n \) statistics terms.

**Lemma 4.** If \( Q \) is a query graph representing a clique typed multi-way spatial join of \( n \) relations and \( r_i \) is a node in \( Q \) representing a relation, the following 4 statistics per \( r_i \) \((1 \leq i \leq n)\) are needed to estimate the result size of \( Q \):

\[
N_{r_i}, L_{r_i,x}, L_{r_i,y}, L_{r_i,xy}.
\]

**Proof.** As in the case of the proof of Lemma 3, the selectivity \( S \) in Eq. (12) is a multiplication of the two addition terms, and each addition term is an expression per axis. For \( x \)-axis addition term, a production term has an object term \( s_{r_i,k_i,x} \) or 1 for \( r_i \). Similarly for \( y \)-axis, \( s_{r_i,k_i,y} \) or 1. If we multiply two addition terms, a production term in the resulting expression has \( s_{r_i,k_i,x} \ast s_{r_i,k_i,y} \) or \( s_{r_i,k_i,x} \ast s_{r_i,k_i,y} \) or 1 for \( r_i \). Since \( \text{Size}(Q) \) is \( \sum_{k=1}^{N_{r_1}} \cdots \sum_{k=1}^{N_{r_n}} S \), the object terms in \( S \) are converted to the relevant statistics terms in \( \text{Size}(Q) \). Therefore \( L_{r_i,xy}, L_{r_i,x}, L_{r_i,y} \) and \( N_{r_i} \) for \( r_i \) in \( \text{Size}(Q) \) are needed.

From the above two lemmas, we are able to know that the complexity of the result size estimation for the clique typed query is much simpler than that for the tree typed query. In Lemma 4, if the \( x \)- and \( y \)-length distributions are independent, we do not need the \( L_{r_i,xy} \) statistics because it is calculated by \( L_{r_i,x} \ast L_{r_i,y}/N_{r_i} \). And in such a case, Eqs. (11) and (12) give the same result.

### 4. Conclusions

Recent research introduced formulas, i.e., Eqs. (5) and (11), for estimating the result size of the two kinds of the multi-way spatial join, tree and clique, under the assumption of the uniform placement distribution. However, the formulas are applied only for special cases of data areas. Eq. (5) is satisfied only in the case that the \( x \)- and \( y \)-lengths of all data per relation are equal, and Eq. (11) in the case that the \( x \)- and \( y \)-length distributions are independent.

In this paper, we extended the formulas to the arbitrary area of spatial data (Eqs. (6) and (12)). The resulting formulas consist of object terms. In order for a query optimizer to estimate the query result size, we transformed the formulas of the object terms into those of the statistics terms. Then we analyzed the time and temporary memory space complexity of the formulas represented by the statistics terms (Lemmas 1 and 3). Finally and most importantly we derived the number and sort of the statistics which are required for the estimation of the multi-way join result size (Lemmas 2 and 4). Lemmas 2 and 4 are also related to the disk and memory space complexity. According to the result of the analysis, the complexity and the number of statistics for the tree typed multi-way spatial join are much more than those for the 2-way join and the clique typed multi-way join.

Although we considered rectangles only in the two-dimensional space, the result of this paper can be easily extendable to the \( n \)-dimensional space. We brute-forcely transformed the expressions of object terms for the tree typed multi-way spatial join into those of statistics terms. As a result, the resulting expressions have many duplicated statistics terms. If we factor out the common statistics terms, we will be able to decrease the time and memory complexity for calculating the query result size. Therefore, as future work, we will devise an efficient factoring algorithm to reduce the complexity for the tree typed multi-way spatial join. According to Lemma 2, in order to estimate the result size of the tree typed multi-way spatial join, we need many statistics per relation. Therefore we will also develop a method to calculate an approximated result size using only a small number of statistics for the tree typed multi-way spatial join.
Acknowledgements

We would like to thank Professor Kazuo Iwama, editor, for his help and the anonymous referees for their valuable comments. This research was supported by the Software Technology Enhancement Program 2000 of the Ministry of Science and Technology of Korea and the Basic Research Program (grant number 99-2-315-001-3) of the Korea Science and Engineering Foundation.

References