Narrative for Paper “Query Processing Using Distance Oracles for Spatial Networks”

Xiaofei Zhao
& Rahul Saladi

1. Problem statement:
The computation of distance along a network is critical for most of the spatial operations. However, it is not easy to do that in an efficient way. You can choose to store the \( n^2 \) distances between all pairs, which are not a good option because of the large space requirement. Therefore, an efficient way to do the real-time execution for queries on transportation networks is needed. For spatial network, it is worth considering if we can gain a significant improvement in the efficiency by losing an affordable accuracy. Especially for the sufficiently large network, it will be unnecessary and unpractical to waiting for the queries to get the exact distance between two vertices.

2. The major contributions of the paper:
First, adapting well-separated pair decomposition technique to a spatial network. Second, using distortion spectrum to improve the resulting error \( \varepsilon \) of the distance oracle, which is used to bound the network distance. Third, building a linear size oracle for finding shortest distance. Specially of size \( O\left(\frac{n}{\varepsilon^d}\right) \)
Four, formulation of simple SQL constructs for answering spatial queries. The operations, such as region search, k-nearest neighbor, and spatial joins, can all be seamlessly integrated into RDBMS by using SQL language.
The paper also talks about some query optimization strategies to adopt the \( \varepsilon \)-approximate distance oracle.

The most significant contributions of this paper are the first one and the second one. The main idea of this paper is expending a little more time and losing an affordable accuracy to decrease the storage requirements.
Many previous works have done based on the search on graphs, which is hard to implement by using the relational operators of database. So in this paper, by introducing well-separated pair decomposition technique to spatial network, we can apply approximate distance oracle to estimate the spatial distance. The experiments results prove that the loss of time and accuracy is worthy compared with what we gain.
The construction of the oracle enables us to capture the \( n^2 \) network distances in a spatial network using only \( O(n) \) space. What’s more significant, it can be seamlessly integrated into a relational databases system, indexed using a traditional indexing scheme, and used in complicated query processing scenarios without making any significant changes to existing databases system.

3. Key concept:
WSP (well-separated pairs): Two sets of points of A and B are said to be well separated if the separation distance between A and B is at least $s \times r$, where $s > 0$ is a separation factor and $r$ is the larger diameter of the two sets.

Network distance $d_G(u,v)$ and spatial distance $d_S(u,v)$: Network distance is the distance along the shortest path between $u$ and $v$ in the spatial network. Spatial distance is a function of the position of the vertices $u$ and $v$ on the embedding plane. These two concepts are used to define the minimum and maximum distortion of $G$.

Approximate distance oracle: Given a spatial network $G(V,E)$ with minimum $\gamma_L$ and maximum $\gamma_H$ distortions, $S_\varepsilon = \{ \gamma_L, \gamma_H \}$ is an $\varepsilon$-approximate distance oracle of unit size such that the approximate network distance between a source vertex $u$ and a destination vertex $v$ is given by $S_\varepsilon(u,v) = (\gamma_L + \gamma_H)/2d_s(u,v)$, where $\varepsilon \leq (\gamma_L - \gamma_H) / (\gamma_L + \gamma_H)$, which can be computed in $O(1)$ time.

One simple exercise may be: (1) for a WSPD, is it required that the symmetric pairs of a set must also appear in the decomposition? The answer is no. (2) It’s easy to see that computing the minimum and maximum distortions for the whole spatial network is not a good choice, which could cause large resulting error. Is there any easy way to improve this? The answer is capturing the distortion spectrum of a spatial network.

4. Validation methodology:

This paper uses experiments to evaluate the query processing method.

First, it shows that $O(1)$-size oracle cannot provide a reasonable answer because the resulting error $\varepsilon \approx 1$ is very large. But for an $O(n)$-size oracle, the techniques of this paper are proved to be applicable to large road networks. Then the paper compared the maximum, average, and the standard deviation of the error according to do 100,000 $\varepsilon$-approximate distance queries on the actual network distance. The errors is low enough to prove that, in practice, the quality of the answers provided by the oracle is very close to the exact network distance. After that, the paper compares the performance of the region search query and the resulting speedup of the proposed techniques. The result shows that the resulting errors are quite low, and the optimized algorithm is much faster than the un-optimized algorithm.

The second set of experiments proves the applicability of distance oracles on the k-nearest neighbor finding algorithm. At last, the experiment also verifies the query optimization can do a large savings in time and distance computations.

Strengths: It is reasonable to use the experiment as the validation method. Because there are three critical measure factors (time, error, space) in this paper need to be tested in order to prove the applicability of the paper. The experimental results could provide a comprehensive comparison between the classic queries and the optimized queries.
Weakness: the variables of the experiment are tightly controlled, which may not be true in realistic.

5. Assumptions:
   1. The distortion between the network and spatial distances between any source and destination vertices in spatial network G is bounded both from above and below.
   2. Given a point set R in a d-dimensional space, assume R is contained in a unit $[0,1]^d$ d-dimensional hypercube. The discussion in this paper is restricted to a spatial network that is embedded in a two-dimensional space where the objects are points specified by their location in terms of latitude and longitude. Because of this assumption, when applying the WSPD on the set of vertices V on a spatial network, the discussion does not need to resort to the path-compressed quadtree while still using regular decomposition.
   3. Given a set of source vertices A and a set of destination vertices B, if A and B are sufficiently far away from each other, which means there are no proximity between A and B, then the network distance between vertices in A and those vertices in B will be similar and can be represented by a single value.

   These assumptions are reasonable in the context of this paper. But they also restrict the applicability of the methods. For instance, it can only be applicable in a sufficiently large network.

6. If I were to rewrite this paper today, most of the ideas in this paper would be preserved, but just remove some constraints.
   1. As we can see that the discussion of this paper is on a fairly large spatial network, so it will be helpful if there is any way to weaken this constraint. Because most of the queries would be local-based.
   2. The paper is based on a 2-dimensional network. A possible extension of the paper is to extend the spatial data to 3D, which may be helpful for some related industry, such as airways.