

Research Statement

Kyriakos MOURATIDIS
School of Information Systems, Singapore Management University
Tel: (65) 6828-0649; Email: kyriakos@smu.edu.sg

January 2012

My research focuses on Spatiotemporal Databases and Location-based Services, and their bridging with Mobile Computing and Data Stream Processing. I am also working on Location Privacy, and Query Verification in Outsourced Databases. Specifically, most of my published results fall under one of the following four topics.

1) Continuous Query Monitoring

Traditional database systems are designed to answer transient, *snapshot* queries over persistent data. However, the evolution of wireless communications, positioning devices (e.g., GPS) and sensor technologies has recently given rise to a new data processing model. In this model, multiple long-running queries require continuous evaluation as the data dynamically change. These queries are called *continuous* (or *standing*), and they arise in location-based services (e.g., “keep me updated about who are the 10 SMU students that are closest to my location as I walk along Orchard road”), network traffic monitoring (e.g., “monitor the 100 users that cause the highest network overhead”), online decision support systems (e.g., “continuously report the 5 most interesting stocks according to my investment criteria”), and many other domains.

In [1], [2], [3] and [4] we consider k -NN monitoring for various settings and with different performance goals. A standing k nearest neighbor (k -NN) query continuously reports the k among a set of moving objects that lie closest to a (potentially moving) query point. [1] describes a method targeted to Euclidean spaces that aims at minimizing the computational overhead for centrally processing multiple queries. It achieves low running time by handling location updates only from objects that fall in the vicinity of some query. In [2] we tackle the same problem, but target at reducing the communication cost between the central processing server and the data objects. We present a threshold-based algorithm that exploits the computational capabilities of the objects to achieve this goal. [1] and [2] assume the Euclidean distance metric. In most real-world scenarios, however, the data objects and the users (queries) move in road networks, and the distance between them is defined as the length of the shortest path connecting them. Motivated by this fact, in [3] we propose efficient k -NN monitoring methods for road networks. In addition to object and query movement, our algorithms take into account changes in the network, e.g., edge weight updates due to varying traffic conditions. All the aforementioned techniques are designed for *update streams*, where the central server processes location updates from a particular set of queries and objects. On the other hand, [4] considers the *sliding window* model where data tuples (spatial points) continuously stream in the server. Each arriving tuple is considered as a new, individual datum, and processing is restricted to the most recent ones only. [4] describes time- and space-efficient algorithms for continuous k -NN queries in this setting. All the above methods consider centralized processing; in [5], instead, we focus on a distributed environment, and specifically that of *wireless broadcasting*. In this setting, the data are repeatedly broadcast by the server, interleaved with some indexing information. Clients (mobile devices) tune in the broadcast channel and monitor their k -NNs locally, without contacting the server. The objective here is to minimize the number of data packets received by the clients, in order to preserve their battery resources.

Monitoring queries of a non-spatial nature is also important. In [6] we continuously evaluate multiple top- k queries. Each such query specifies a preference function f and requests the k tuples in a sliding window that have the highest scores according to f . Our objective is to monitor these queries with a low computational cost at the processing server. In [7] we address the processing of continuous text queries over streams of documents. The challenge here is that documents essentially define a high-dimensional space, where spatial indexing and geometric reasoning fail (due to the dimensionality curse). Thus, we use a dynamic adaptation of the well-established inverted index (which is specific to text documents), and propose incremental algorithms that exploit this data structure.

2) Spatial Optimization

This category includes problems that (i) arise in resource allocation applications, (ii) have a spatial nature (i.e., their optimization criteria involve the notion of distance), and (iii) are of a large scale (thus, typically requiring disk-based storage). Although such problems have been dealt with in the operations research literature, the solutions proposed there are targeted to in-memory processing and are suitable for small size datasets. The below described work enables efficient processing in cases where scale is several times (or orders) larger than considered before.

In [8] we introduce and process *aggregate nearest neighbor* (ANN) queries. An ANN query retrieves the k data objects (from a disk-resident spatial dataset) with the smallest aggregate distance from a set of query points. The aggregate distance could be *sum*, *min*, or *max*. For example a *sum*-ANN query could find the facility (e.g., restaurant) where a set of users can meet by travelling the smallest possible total distance. In [9] we solve the *k-medoid* problem, assuming the existence of a spatial index on the input dataset. Suppose that a franchise plans to open k branches in a city, so that the average distance from each residential block to the closest branch is minimized. This is an instance of the *k-medoids* problem, where residential blocks constitute the input dataset and the k branch locations correspond to the medoids. In [10] we solve the optimal matching problem (one of the oldest problems in operations research) in the context of large spatial databases. Consider a set of *customers* (e.g., WiFi receivers) and a set of *service providers* (e.g., wireless access points), where each provider has a *capacity* and the quality of service offered to its customers is anti-proportional to their distance. Our task is to compute a matching between the two sets such that (i) the maximum possible number of customers are served, and (ii) the sum of Euclidean distances within the assigned provider-customer pairs is minimized. Although max-flow algorithms exist for this problem, they are inapplicable to medium or large scale problems, because their memory requirements exceed several times the main memory size of a typical server. Motivated by this fact, we propose efficient algorithms for optimal assignment that employ novel pruning strategies, based on the geometric properties of the problem. In [11] we extend the above study, proposing new algorithms that rely again on a mixture of operations research algorithms and spatial pruning. Moreover, we consider incremental assignment maintenance in the presence of customer updates. In [12] we study a related problem with two important differences: (i) each service provider has a *coverage region* (i.e., it can only serve customers within a specific area), and (ii) the customers move frequently and arbitrarily. The task of the processing server in this scenario is to continuously report the optimal assignment (subject to the customers' most recent locations) in an online fashion. Unlike [10] and [11], the additional constraints imposed by the coverage regions allow for purely geometric criteria to reduce the running time, while treating the underlying operations research tools as black boxes. An interesting feature in [12] is that our methodology allows for parallelization (and thus even faster processing), which is a very desirable feature given the recent availability of inexpensive multi-core CPUs in the hardware industry.

In [13] we consider a *stable marriage* problem (yet another traditional resource allocation topic) where our task is to produce a fair assignment between a set of user preference queries and a set of available facilities. As an example consider an internship assignment system, where at the end of each academic year, interested university students search and apply for available positions, based on their preferences

(e.g., nature of the job, salary, office location, etc). Although this is not a spatial problem per se, when the facility attributes are few (i.e., the dimensionality is low) geometric reasoning can be used to significantly reduce the processing time (CPU and I/O cost).

In [14] we study a problem which falls in the intersection of spatial databases and multi-criteria decision making. It is the first work in the area of road network databases that takes into account the co-existence of multiple distance notions in transportation decisions. For example, the different costs of a road segment could be its Euclidean length, the driving time, the walking time, possible toll fee, etc. The relative significance of these proximity notions may vary from user to user, yielding different location-based preferences over the facilities reachable via the network. In [14] we formalize preference queries over facilities located in a multi-cost road network, and design algorithms for their efficient processing.

3) Location Privacy

The increasing trend of embedding positioning systems (e.g., GPS) in mobile devices facilitates the widespread use of location-based services. For such applications, the privacy and confidentiality issues are of paramount importance. Existing techniques, like encryption, safeguard the communication channel from eavesdroppers. Nevertheless, the queries themselves may disclose the position, identity and habits of the user. In [15] we propose location obfuscation methods for preserving the anonymity of the users (in Euclidean space), and design processing methods to evaluate spatial queries from obfuscated locations. On the other hand, many location-based services involve users and objects that lie in a road network; in this context, proximity is determined by network distance, which cannot be captured by [15] or other techniques designed for Euclidean spaces. In [16] we propose the first framework for anonymous query processing in road networks. In [17] we identify an ambiguity in the privacy requirements traditionally adopted in the spatial anonymity literature, with serious implications in performance. Our investigation reveals the existence of a "hidden" dimension in this area (that of information leakage), which we isolate, formalize, and measure. In addition to clarifying a fundamental ambiguity, we show how existing methods may build on our analysis to improve performance without violating anonymity.

Besides location obfuscation, spatial anonymity principles also apply to the privacy-preserving publishing of sensitive information, such as medical records. In [18] we make use of such techniques in conjunction with an important observation on the role of publically available information, and enhance the usability of published data (i.e., reduce information loss), without compromising the privacy of individuals.

4) Database Authentication

In the outsourced database model, a data owner publishes his/her database through a third-party server; i.e., the server is responsible for hosting the data and answering user queries on behalf of the owner. Since the server may not be trusted, or may be compromised, users need a means to verify that answers received are both authentic and complete, i.e., that the returned data have not been tampered with, and that no qualifying results were omitted. In [19] we propose verification methods for static and dynamic databases (applicable to one-dimensional and spatial data) that achieve fast query processing, and low storage and communication overheads. In [20] we revisit a previously dismissed authentication model, and show its viability with modern hardware. Furthermore, we propose efficient techniques for the verification of join results, a problem for which no practical solution existed. Also, we address the issue of result freshness, preventing the server from suppressing database updates (i.e., hiding them from the users). In [21] we authenticate collections of documents and design a framework that verifies the results of text search engines on these documents. In [22] we propose methods for shortest path verification in outsourced graphs, such as transportation networks.

Ongoing and Future Work

My ongoing research centers on the above axes, and considers similar or related problems in domains where they have not been addressed before. My main long-term goal is to integrate spatial processing techniques into emerging data models (such as data streams, wireless broadcasting, outsourced databases, etc), and to address potential user privacy and result authenticity issues that arise thereof.

References

- [1] K. Mouratidis, M. Hadjieleftheriou, D. Papadias: Conceptual Partitioning: An Efficient Method for Continuous Nearest Neighbor Monitoring. *ACM Conference on Management of Data (SIGMOD)*, 2005.
- [2] K. Mouratidis, D. Papadias, S. Bakiras, Y. Tao: A Threshold-based Algorithm for Continuous Monitoring of k Nearest Neighbors. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 17(11), 1451-1464, 2005.
- [3] K. Mouratidis, M. Yiu, D. Papadias, N. Mamoulis: Continuous Nearest Neighbor Monitoring in Road Networks. *Very Large Data Bases Conference (VLDB)*, 2006.
- [4] K. Mouratidis, D. Papadias: Continuous Nearest Neighbor Queries over Sliding Windows. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(6), 789-803, 2007.
- [5] K. Mouratidis, S. Bakiras, D. Papadias: Continuous Monitoring of Spatial Queries in Wireless Broadcast Environments. *IEEE Transactions on Mobile Computing (TMC)*, 8(10), 1297-1311, 2009.
- [6] K. Mouratidis, S. Bakiras, D. Papadias: Continuous Monitoring of Top- k Queries over Sliding Windows. *ACM Conference on Management of Data (SIGMOD)*, 2006.
- [7] K. Mouratidis, H. Pang: Efficient Evaluation of Continuous Text Search Queries. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 23(10), 1469-1482, 2011.
- [8] D. Papadias, Y. Tao, K. Mouratidis, K. Hui: Aggregate Nearest Neighbor Queries in Spatial Databases. *ACM Transactions on Database Systems (TODS)*, 30(2), 529-576, 2005.
- [9] K. Mouratidis, D. Papadias, S. Papadimitriou: Tree-based Partition Querying: A Methodology for Computing Medoids in Large Spatial Datasets. *Very Large Data Bases Journal (VLDBJ)*, 17(4), 923-945, 2008.
- [10] L. U, M. Yiu, K. Mouratidis, N. Mamoulis: Capacity Constrained Assignment in Spatial Databases. *ACM Conference on Management of Data (SIGMOD)*, 2008.
- [11] L. U, K. Mouratidis, M. Yiu, N. Mamoulis: Optimal Matching between Spatial Datasets under Capacity Constraints. *ACM Transactions on Database Systems (TODS)*, 35(2), 2010.
- [12] L. U, K. Mouratidis, N. Mamoulis: Continuous Spatial Assignment of Moving Users. *Very Large Data Bases Journal (VLDBJ)*, 19(2), 141-160, 2010.
- [13] L. U, N. Mamoulis, K. Mouratidis: A Fair Assignment Algorithm for Multiple Preference Queries. *Very Large Data Bases Conference (VLDB)*, 2009.
- [14] K. Mouratidis, Y. Lin, M. Yiu: Preference Queries in Large Multi-Cost Transportation Networks. *IEEE International Conference on Data Engineering (ICDE)*, 2010.
- [15] P. Kalnis, G. Ghinita, K. Mouratidis, D. Papadias: Preserving Location-based Identity Inference in Anonymous Spatial Queries. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(12), 1719-1733, 2007.
- [16] K. Mouratidis, M. Yiu: Anonymous Query Processing in Road Networks. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(1), 2-15, 2010.
- [17] K. Tan, Y. Lin, K. Mouratidis: Spatial Cloaking Revisited: Distinguishing Information Leakage from Anonymity. *International Symposium on Spatial and Temporal Databases (SSTD)*, 2009.
- [18] D. Sacharidis, K. Mouratidis, D. Papadias: k -Anonymity in the Presence of External Databases. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(3), 392-403, 2010.
- [19] K. Mouratidis, D. Sacharidis, H. Pang: Partially Materialized Digest Scheme: An Efficient Verification Method for Outsourced Databases. *Very Large Data Bases Journal (VLDBJ)*, 18(1), 363-381, 2009.
- [20] H. Pang, J. Zhang, K. Mouratidis: Scalable Verification for Outsourced Dynamic Databases. *Very Large Data Bases Conference (VLDB)*, 2009.
- [21] H. Pang, K. Mouratidis: Authenticating the Query Results of Text Search Engines. *Very Large Data Bases Conference (VLDB)*, 2008.
- [22] M. Yiu, Y. Lin, K. Mouratidis: Efficient Verification of Shortest Path Search via Authenticated Hints. *IEEE International Conference on Data Engineering (ICDE)*, 2010.

All papers can be downloaded from: www.mysmu.edu/faculty/kyriakos/publications.html