UNIVERSITI PUTRA MALAYSIA

# *SKYLINE QUERIES ON DATA WITH UNCERTAIN DIMENSIONS FOR EFFICIENT COMPUTATION*

NURUL HUSNA MOHD SAAD

FSKTM 2018 72

# SKYLINE QUERIES ON DATA WITH *UNCERTAIN DIMENSIONS* FOR EFFICIENT COMPUTATION

By

NURUL HUSNA MOHD SAAD

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Doctor of Philosophy

July 2018

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Doctor of Philosophy

**SKYLINE QUERIES ON DATA WITH *UNCERTAIN DIMENSIONS* FOR EFFICIENT COMPUTATION**

By

**NURUL HUSNA MOHD SAAD**

**July 2018**

**Chair : Professor Hamidah Ibrahim, PhD**
**Faculty : Computer Science and Information Technology**

The notion of skyline query is to find a set of objects that is not dominated by any other objects. Skyline query is crucial in multi-criteria decision making applications particularly in applications that generate uncertain data. Although there is a significant amount of research that has been committed for efficient skyline computation, regrettably, existing works lack on how to conduct skyline queries on uncertain data with objects represented as continuous ranges and exact values. By having data with *uncertain dimensions*, the dominance relation among objects with continuous ranges and exact values may not be transitive, thus, causing existing techniques for skyline queries are not applicable. The results of skyline queries are bound to be probabilistic since each object with continuous range is now associated with a probability value of it being a query answer. Furthermore, querying information within a range of search on *uncertain dimensions* proves to be challenging in order to determine objects with continuous ranges that satisfy the range query. Hence, this thesis focuses on efficiently extending skyline query and range skyline query processing to support data with *uncertain dimensions*. We define skyline queries over data with *uncertain dimensions* and present four methods to efficiently answer skyline queries, namely: *distinctive partitioning*, *exact domination*, *range domination*, and *uncertain domination*. We propose a two-phase framework, *SkyQUD*, which integrates these four methods; the first phase employs efficient probability computations which are performed individually on groups of objects with exact values and continuous ranges, respectively. Meanwhile, the second phase employs more complex and expensive computations to perform dominance testing between objects from different groups. The *SkyQUD* framework is responsible to extract the most dominant skyline objects that meet the required threshold value. The threshold value is utilized in order to manage the quality and the size of the skyline objects reported. Next, we extend *SkyQUD* to support skyline with range queries on *uncertain dimensions*, denoted as *SkyQUD-T*. A method, *range pruning*, is proposed and incorporated before the first phase in *SkyQUD* to determine objects that *satisfy* the range query, where it bounds the probability

of each object to a certain threshold value. Both frameworks have been validated through extensive experiments employing real and synthetic datasets. Several independent variables which are *scalability*, *threshold*, *data distributions*, and *dimensionality* are selected to determine their effects on two dependent variables. The effect of manipulating the independent variables is studied on the dependent variables which are *number of pairwise comparisons* and *processing time*. Through theoretical analysis and extensive experiments, we show that *SkyQUD* is able to effectively support skyline queries on data with *uncertain dimensions* and capable of handling large datasets. The performance of *SkyQUD-T* is studied against two naïve algorithms that are developed to reflect the best-case and worst-case scenarios. Results exhibit the evidences of the behaviour of *SkyQUD-T*, where the number of pairwise comparisons performed in *SkyQUD-T* is always within the performance of the aforementioned naïve algorithms.

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

**PERTANYAAN *SKYLINE* PADA DATA DENGAN DIMENSI TIDAK PASTI UNTUK PENGIRAAN EFISIEN**

Oleh

**NURUL HUSNA MOHD SAAD**

**Julai 2018**

**Pengerusi : Professor Hamidah Ibrahim, PhD**
**Fakulti : Sains Komputer dan Teknologi Maklumat**

Pengertian pertanyaan *skyline* adalah untuk mencari satu set objek yang tidak didominasi oleh objek lain. Pertanyaan *skyline* adalah penting dalam aplikasi membuat keputusan pelbagai kriteria terutama dalam aplikasi yang menghasilkan data tidak pasti. Walaupun terdapat sejumlah signifikan penyelidikan yang telah dilakukan untuk pengiraan *skyline* yang efisien, malangnya, kerja yang sedia ada kekurangan dalam bagaimana untuk melakukan pertanyaan *skyline* pada data tidak pasti dengan objek yang diwakili sebagai julat berterusan dan nilai tepat. Dengan mempunyai data dengan dimensi tidak pasti, hubungan dominasi antara objek dengan julat berterusan dan nilai tepat mungkin tidak transitif, oleh itu, menyebabkan teknik sedia ada untuk pertanyaan *skyline* tidak dapat digunakan. Hasil pertanyaan *skyline* terikat untuk menjadi probabilistik kerana setiap objek dengan julat berterusan kini dikaitkan dengan nilai kebarangkalian untuk ia menjadi jawapan pertanyaan. Tambahan pula, pertanyaan maklumat dalam julat pencarian pada dimensi tidak pasti terbukti mencabar untuk menentukan objek dengan julat berterusan yang memenuhi pertanyaan julat. Oleh itu, tesis ini memberi tumpuan terhadap melanjutkan secara efisien pertanyaan *skyline* dan pemprosesan pertanyaan *skyline* julat untuk menyokong data dengan dimensi tidak pasti. Kami mendefinisikan pertanyaan *skyline* bagi data dengan dimensi tidak pasti dan mengemukakan empat kaedah untuk menjawab pertanyaan *skyline* dengan efisien, iaitu: *distinctive partitioning*, *exact domination*, *range domination*, dan *uncertain domination*. Kami mencadangkan dua fasa rangka kerja, *SkyQUD*, yang menggabungkan empat kaedah ini; fasa pertama menggunakan pengiraan kebarangkalian efisien yang dilakukan secara individu pada kumpulan objek dengan nilai tepat dan julat berterusan, masing-masing. Sementara itu, fasa kedua menggunakan pengiraan yang lebih kompleks dan mahal untuk melakukan pengujian dominasi antara objek dari kumpulan yang berlainan. Rangka kerja *SkyQUD* bertanggungjawab untuk mengekstrak objek *skyline* yang paling dominan yang memenuhi nilai ambang yang diperlukan. Nilai ambang digunakan untuk menguruskan kualiti dan saiz objek *skyline* yang dilaporkan. Seterusnya, kami melanjutkan *SkyQUD* untuk

menyokong *skyline* dengan julat pertanyaan pada dimensi tidak pasti, yang ditandakan sebagai *SkyQUD-T*. Satu kaedah, *range pruning*, dicadangkan dan digabungkan sebelum fasa pertama di dalam *SkyQUD* untuk menentukan objek yang memenuhi julat pertanyaan, di mana ia mengikat kebarangkalian setiap objek kepada nilai ambang tertentu. Kedua-dua rangka kerja telah disahkan melalui eksperimen yang menyeluruh menggunakan set data sebenar dan sintetik. Beberapa pembolehubah bebas, iaitu skalabiliti, ambang, pengagihan data, dan dimensi dipilih untuk menentukan kesannya kepada dua pembolehubah bergantung. Kesan memanipulasi pembolehubah bebas dikaji pada pembolehubah bergantung tersebut, iaitu bilangan perbandingan pasangan dan masa pemprosesan. Melalui analisis teori dan eksperimen yang menyeluruh, kami menunjukkan bahawa *SkyQUD* dapat menyokong pertanyaan *skyline* secara efektif pada data dengan dimensi tidak pasti dan mampu menangani set data besar. Prestasi *SkyQUD-T* dikaji terhadap dua algoritma naif yang dibangunkan untuk mencerminkan senario terbaik dan terburuk. Hasil mempamerkan bukti tingkah laku *SkyQUD-T*, di mana bilangan perbandingan pasangan yang dilakukan di *SkyQUD-T* sentiasa berada dalam prestasi algoritma naif yang dinyatakan di atas.

# ACKNOWLEDGEMENTS

First and foremost, I am eternally thankful to Allah for his blessings, strength, and perseverance bestowed upon me, enabling me to complete this research successfully. I wish to express my deepest gratitude to my advisor and mentor Professor Dr. Hamidah Ibrahim who worked closely with me, taught me countless valuable research skills without reserve, kept an eye on the progress of my work, and guiding me through the PhD program. She always found time for discussing research, reading my texts, and giving good advices; without her this thesis would not have been possible. I really appreciate her help to improve the quality of my thesis. I would have been lost without her. I am also very thankful to my advisory committee, Associate Professor Dr. Fatimah Sidi, Associate Professor Dr. Razali Yaakob, and Associate Professor Dr. Azmi Jaafar, for their insightful comments, encouragement, sound advice, and lots of good ideas. Having the opportunity of working together with them was a grateful learning experience that made me a better researcher. I am indebted to the Ministry of Higher Education (MOHE) that has granted me the MyBrain15 scholarship and to University Putra Malaysia for the opportunity of studying the PhD program. Thanks for this amazing opportunity. I would also like to thank my family and beloved friends that have always encouraged me. Thanks for the talks during the coffee, for the dinners we had together, and for all encouragement provided during these years. I dedicate a special acknowledge to my parents that started all of this. They have dedicated part of their lives for providing education to me, and I owe this achievement to them. Finally, my thanks go to all the people who have directly or indirectly supported me in completing this research work.

I certify that a Thesis Examination Committee has met on 12 July 2018 to conduct the final examination of Nurul Husna binti Mohd Saad on her thesis entitled "Skyline Queries on Data with *Uncertain Dimensions* for Efficient Computation" in accordance with the Universities and University Colleges Act 1971 and the Constitution of the Universiti Putra Malaysia [P.U.(A) 106] 15 March 1998. The Committee recommends that the student be awarded the Doctor of Philosophy.

Members of the Thesis Examination Committee were as follows:

**Md Nasir bin Sulaiman, PhD**
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

**Rohaya binti Latip, PhD**
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

**Lilly Suriani Affendey, PhD**
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

**Dominique Laurent, PhD**
Professor
Cergy-Pontoise University
France
(External Examiner)

**RUSLI HAJI ABDULLAH, PhD**
Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date: 30 July 2018

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy. The members of the Supervisory Committee were as follows:

**Hamidah Ibrahim, PhD**
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

**Fatimah Sidi, PhD**
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

**Razali Yaakob, PhD**
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

**ROBIAH BINTI YUNUS, PhD**
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

**Declaration by graduate student**

I hereby confirm that:

- this thesis is my original work;
- quotations, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research) Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: ________________________ Date: __________________

Name and Matric No.: _______________________________________

**Declaration by Members of Supervisory Committee**

This is to confirm that:

- the research conducted and the writing of this thesis was under our supervision;
- supervision responsibilities as stated in the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: ______________________________
Name of Chairman of Supervisory Committee: ______________________________

Signature: ______________________________
Name of Member of Supervisory Committee: ______________________________

Signature: ______________________________
Name of Member of Supervisory Committee: ______________________________

TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

Traditional queries in database systems are usually represented by a set of predicates where the results return are expected to meet the requirements of the predicates precisely, and consequently the product of these queries would either be a set of exact results or an empty set. These traditional queries are effective when there are no expectations of comparisons of values between different dimensions. Nevertheless, employing queries with the aforementioned hard constraints in applications where users want to retrieve the best results but at the same time have conflicting preference for each dimension is impractical as the resulting queries will be an empty set.

Therefore, recent approaches towards intuitive information systems and the integration of user preferences have led to skyline queries. The skyline retrieval paradigm has received a lot of attention since a decade ago as it proved especially useful for query personalization (Kießling, 2002; Koutrika and Ioannidis, 2004). Skyline queries introduce the notion of dominated objects under Pareto optimality. In a database consisting of multi-dimensional objects, the skyline queries performed on the underlying database would return a set of objects that is the best trade-offs between all dimensions involved.

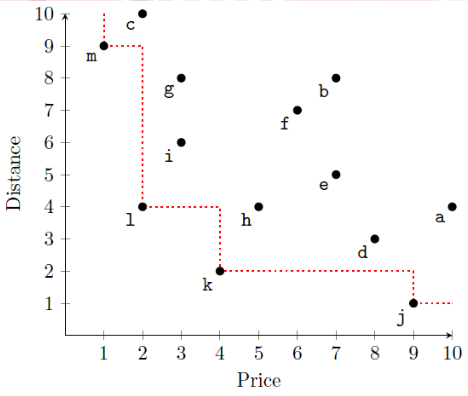


**Figure 1.1: Skyline example**

For example, consider a database that contains information on hotel’s room rate per night and distance to the beach. Figure 1.1 shows each record of the database that is represented as a point in a data space consisting of those two dimensions. A user may posed a query such as “*find me hotels that are as* *cheap as possible and as close as possible to the beach*". Now, the query itself can be understood differently depending on the user, the user might have wanted the cheapest price a hotel can offer with the hotel being not really close to the beach, or the user would have been satisfied with paying extra cost for the hotel as long as it is as near as possible to the beach. Furthermore, it is almost impossible to answer the query of how much cheaper and nearer a hotel should be. In this case, the traditional query is roughly written in SQL syntax as follows:

```
SELECT          *
FROM            Hotels
WHERE           Price MIN,
                Distance MIN;
```

Querying the above query on the database in Figure 1.1 would return an empty result set as there is no object that matches with the requirements of the query. To remedy this, the database community has proposed to extend the SQL's SELECT statement by injecting skyline functionality as follows (Börzsönyi et al., 2001):

```
SELECT          *
FROM            Hotels
SKYLINE OF      Price MIN,
                Distance MIN;
```

The skyline technique identifies a set of objects, *S*, in such a way that they are not dominated by the other objects in the dataset. In other words, an object $v$ is *preferred* over another object $w$ if and only if $v$ is better than $w$ strictly in at least one dimension and $v$ is better than or equal to $w$ in all other $d$ dimensions. With the above skyline query, it will retrieve, as illustrated by the dotted red line in Figure 1.1, all hotels that are either cheap or near to the beach (or both, if possible) without there being other hotels that are cheaper and nearer to the beach. Therefore, it is ideal to find skyline objects to aid in narrowing down the candidates.

## 1.2 Motivation

Uncertainty in data is not surprising as it can arise in a variety of scenarios, such as when attempting to preserve privacy thus data modification is performed, missing data are approximated, collecting trajectories data which are spatiotemporal in nature, and inadequate equipment for measurement in sensor network. One of the many scenarios can be seen when extracting web data to perform analysis on it. As more and more data become accessible via web servers, information has to be efficiently retrieved from these databases. In extracting web data, uncertainty and incompleteness could not be avoided and no matter what the cause of uncertainty is, it is essential to be able to handle the uncertainty of data.

For example, a property investor in Vancouver would like to do analysis for insights into the rental market. In order to do the analysis, a search on *Rent.com* web page (*www.rent.com*) for "Vancouver" is performed and the search results are extracted. For illustration, Figure 1.2 shows the result of extracting web data from rental listings. Data extraction has caused uncertainty in the dimensions, for instance, the rent dimension, where some of the rent values are uncertain; they are presented as either an exact value or a continuous range. Attempting to perform skyline queries on this kind of data would be challenging due to the dissimilarities nature of data – we could be attempting to compare apples with oranges.

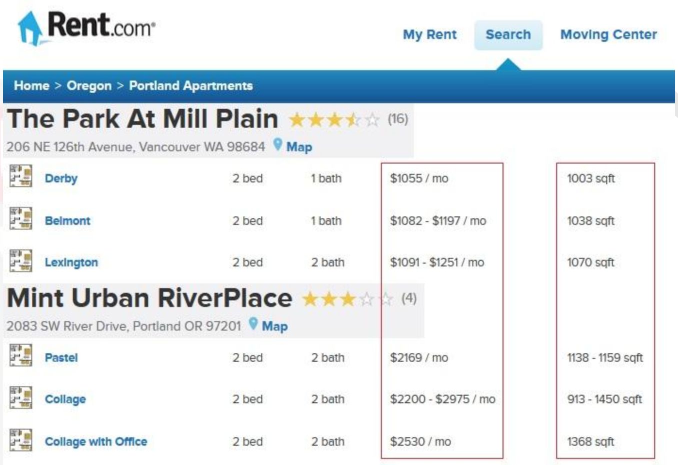


**Figure 1.2: Extraction of web data on rentals from *Rent.com***

One naïve approach to avoid the uncertainty is to use an automated translator or perform digital curation to reformat data from one representation to another. This approach has an obvious drawback, in that the transformation of data does not guarantee that the combined, transformed data are meaningful, and thus, performing skyline queries on the underlying data would as well incur inaccurate result of skyline objects.

There are a plethora of scenarios, other than the aforementioned scenario, where uncertainty in data is inherent and inevitable due to the rapidly evolving technologies and the growing amount of uncertain data generated and accumulated daily. Conversely, it is undeniable how imperative it is to have data analysis techniques, such as the preference evaluation queries, that are able to accommodate any kind of uncertainty in data. Uncertainty in data has led to several new challenges to the database community, which has led to the need for a simple yet efficient algorithm that is able to perform advanced analysis on uncertain data. It remains an open problem at large on how to conduct advanced analysis, particularly skyline analysis, which will be the focus in this thesis, when there are uncertainties in data.

## 1.3 Research Problem

As we have elucidated in Section 1.1, skyline queries have become one of the most popular and frequently used preference queries in the attempt towards intuitive information systems. As with all approaches, skyline queries are not without its own limitations, be it on certain or uncertain data, in that the complexity of skyline computation is greatly influenced by the number of dimensions as well as the size of the dataset. The search space of skyline queries is also highly affected by the number of dimensions in the dataset. This is because as the number of dimensions increases, the size of the searching space will increase extensively as it will become more difficult for objects to be dominated (Yiu and Mamoulis, 2007; Siddique and Morimoto, 2009). Therefore, most of the existing works (Börzsönyi et al., 2001; Tan et al., 2001; Kossmann et al., 2002; Papadias et al., 2003; Chomicki et al., 2003; Papadias et al., 2003; Godfrey et al., 2005; Pei et al., 2005; Yuan et al., 2005; Wu et al., 2006; Tao et al., 2006; Chan et al., 2006; Dellis and Seeger, 2007; Cui et al., 2008) on skyline queries have attempted to solve these limitations by minimizing the searching space as small as possible and reducing the processing cost of finding skylines. The searching space in skyline queries normally is determined by the number of pairwise comparisons that needs to be performed between objects in order to identify skyline objects. To summarise, the larger the size of the searching space, the higher the number of pairwise comparisons, which in turn will result in an expensive cost of processing skyline queries.

Other than data that are certain, skyline query analysis is found to be important on data that are uncertain as well, where certainty and preciseness are lacking in the data. Uncertainty in data is inherent and unavoidable, which is normally caused by data randomness and incompleteness, noises during measuring, delayed during data updates in sensors, or limited knowledge. In general, solving the limitations of skyline queries is a challenge on itself on data that are certain, let alone on data that are uncertain. Attempting to perform skyline query analysis on uncertain data proves to be another major challenge as the data are now incorporated with uncertainties which affects the skyline analysis as well. Hence, the main focus of this thesis is on solving the limitations of skyline queries on uncertain data.

Following the above research problem, we have identified two major challenges that we address in this thesis, which focus on the analysis and computation of skyline queries on uncertain data.

**Challenge 1: Efficient Computation of Skyline Queries on Uncertain Data**

When dealing with uncertainty in data, it does not seem desirable to completely eradicate the uncertain values as it may lead to inaccurate or incomplete query results. Generally, there are two types of model when dealing with uncertainty in data, namely: discrete and continuous. In discrete uncertainty model, each object has a finite set of possible values, called instances, each of which is associated with a probability distribution. On the contrary, in continuous uncertainty model, each object is represented by continuous ranges, which is associated with a probability density function capturing the likelihood of possible values.

Although there has been research done on handling uncertainty in skyline query processing, they hold onto the assumption, with respect to the above uncertainty model, that either (i) uncertainty in data is caused by multiple existences of instances that represent an object (Pei et al., 2007; Yong et al., 2008; Böhm et al., 2009; and Atallah and Qi, 2009), or (ii) the occurrence of uncertainty in a dimension would mean all values under that dimension is represented as a continuous range (Khalefa et al., 2010). Unfortunately, in situations where values in a dimension are presented as both exact values and continuous ranges, such assumptions do not capture well the nature of data uncertainty in each dimension, where imperfections of data caused by uncertain values are inherent in today's real world application.

Given a set of objects with uncertain data, where the underlying uncertainty in this thesis means that in a dimension (attribute), the objects may be represented as an exact value and/or a continuous range. We refer to this uncertainty in data as *uncertain dimensions*. How can we determine the dominance relation between objects that have different representations of values in each dimension? Given the underlying uncertainty in the data, which object shall be the skyline objects on those uncertain data? Can efficient methods be developed to efficiently compute skyline probabilities when encountering those uncertain data? To the best of our knowledge, the study by Li et al. (2012) is the only work that has endeavored to answer the above questions, with respect to the motivating example described in Section 1.2, by presenting an algorithm, named *BBIS*, which performs the dominance testing between objects by comparing their median values (center points) and is indexed by an R*-tree structure. Despite the contributions made by Li et al. (2012), *BBIS* has its shortcoming when performing on high dimensionality dataset. This is largely due to the poor performance of the R*-tree index structure as it is well known that R*-tree could not adequately indexed large data of more than 5 dimensions (Berchtold et al., 1996; Papadias et al., 2005; Li et al., 2012).

**Challenge 2: Efficient Computation of Skyline with Range Queries on Uncertain Data**

As have been discussed in previous Challenge 1, it is irrefutable that computing skyline probabilities on *uncertain dimensions* is much more complicated than computing skylines on conventional data. It becomes even more intricate particularly when a search region is being queried on the *uncertain dimensions* and skyline objects are expected to be reported within the search region. With the notion that in an *uncertain dimension* an object may be represented by a continuous range, when taking into consideration a range query on the underlying dimension, we have to consider the many possible ways that an object may intersect with the range query and other objects. Granted that range queries are not some new issues faced by the database community, especially when analysing skyline objects within a range query. Cases in point are the works by Papadias et al. (2005), Jiang and Pei (2009), Wang et al. (2011), Rahul and Janardan (2012), Lin et al. (2013), Chester et al. (2014), Mortensen et al. (2015), and He et al. (2016), where they introduced several algorithms to process skyline with range queries on traditional data where there are no uncertainties involved. Despite having a plethora of studies focusing on skyline query and its variant, thus far to the best of our knowledge, there is no work accomplishes on skyline with range query regarding uncertain data.

Given a range query, how can we determine to either accept or reject objects that intersect with the boundary of a range query? Can efficient methods be developed to efficiently compute skyline probabilities within a range query when dealing with a set of objects that has different representations of values in each dimension? Even though the first question has been covered in the area of range queries, for instance the work by Wolfson et al. (1999), yet the focus of the work is mainly in the context of tracking a moving object, and it does not however cover on the techniques for evaluating skyline together with the range queries. Therefore, in our work we attempt to solve the issue of analysing skyline on uncertain data with respect to the range queries.

## 1.4 Research Objectives

The main aim of this research work is to propose a framework that is able to perform skyline queries computation on data with *uncertain dimensions*. In order to achieve the aim, we present the following objectives with respect to the challenges posed in Section 1.3:

1) To propose an efficient algorithm that is able to determine skyline objects on uncertain data, with the aim of reducing the number of pairwise comparisons by eliminating the dominated objects as early as possible in order to avoid unnecessary probability computations.
2) To propose an extension to the aforementioned algorithm that is able to retrieve skyline objects that *satisfy* a given range query, which aids in reducing the searching space in identifying skylines.

## 1.5 Research Scope

The scope of this research work is outlined in the following points:

- The type of query that is considered in this research study is the preference queries, in which case is limited to skyline queries as it is the most frequently applied technique in most of multi-criteria decision making applications (Börzsönyi, et al., 2001; Tan et al., 2001; Kossmann et al., 2002; Chomicki et al., 2003; Yuan et al., 2005; Pei et al., 2005; Godfrey et al., 2005; Bartolini et al., 2006; Pei et al., 2007; Khalefa et al., 2010; Li et al., 2012).

- This research study used the relational data model as it is the most dominant model among the conventional models (Yuan et al., 2005; Pei et al., 2007; Khalefa et al., 2008; Khalefa et al., 2010; Li et al., 2012).
- This research work concentrates on two types of database, namely: synthetic and real databases. The synthetic datasets are of the independent, correlated, and anti-correlated distributions, while the real dataset is from the National Basketball Association (NBA) statistics. These are the most popular types of database that have been used in this area (Börzsönyi, et al., 2001; Pei et al., 2007; Yong et al., 2008; Böhm et al., 2009; Atallah and Qi, 2009; Khalefa et al., 2010; Li et al., 2012).

## 1.6 Contributions

In this thesis, the focus of the study is majorly on the analysis and computation of skyline queries on uncertain data. Despite the contributions made by Pei et al. (2007), Khalefa et al. (2010), and Li et al. (2012) in processing skyline on uncertain data, these works have some drawbacks due to their limited applicability when dealing with large datasets (caused by higher dimensionality and total number of objects in datasets) and massive skyline size (due to high dimensionality in datasets or datasets with anti-correlated distribution (Yiu and Mamoulis, 2007; Siddique and Morimoto, 2009)). Motivated by this fact, we make the following contributions to answer the challenges that are addressed in this thesis.

### Contribution 1

We introduce the concept of *uncertain dimensions* with respect to the motivating example elucidated in previous Section 1.2. Essentially, *uncertain dimensions* are dimensions that contain at least one value that is represented as continuous range. To determine if an object $v$ is preferable over another object $w$ when involving an *uncertain dimension*, we calculate the probability of $v$ to dominate $w$. When both $v$ and $w$ are represented as continuous ranges, we define seven possible types of relations to compute the probability of $v$ to dominate $w$. Alternatively, when $v$ is represented as an exact value while $w$ is represented by a continuous range, we modify the relations defined previously and employed a correction value $\varepsilon$ to accommodate the dominating probability of two objects with different representations. Based on the above relations, we define the dominance relation and skyline on *uncertain dimensions*.

We also utilise a probability threshold value $\tau$ to aid in the pruning process in order to manage the quality and the size of the skyline objects reported. Therefore, the probability of an object to be in the skyline results is the probability that the object is not dominated by any other objects, and that probability is at least $\tau$.

We develop an efficient framework, named *SkyQUD*, for computing skyline probabilities and processing skyline queries on data with *uncertain dimensions*. The *SkyQUD* algorithm determines skyline objects by utilising four optimization methods, namely: *Distinctive Partitioning*, *Exact Domination*, *Range Domination*, and *Uncertain Domination*, with the objective to reduce the number of pairwise comparisons between objects as well as to circumvent unnecessary probability computations, while guaranteeing the probability of each object to be in the final skyline results. This is important as each *uncertain dimension* requires skyline probabilities computations and consequently the larger the number of *uncertain dimensions* in a dataset, the higher the cost of skyline probabilities computations. The *SkyQUD* framework is presented in a general two-phase framework that incorporates the aforementioned four methods to efficiently compute skyline probabilities on *uncertain dimensions*, namely: *harvesting* and *strict selection*. The *harvesting* phase performs a preliminary elimination round to isolate objects that are possible skyline candidates. Massive harvesting of bad or uninteresting objects will mostly occur in this phase. On the other hand, the *strict selection* phase is responsible to extract the candidates of skyline objects that meet the required threshold value $\tau$. Our framework is efficient and scalable as verified by our extensive experimental results in Chapter 6. The experimental results exhibit that our framework is faster than the existing algorithms.

**Contribution 2**

We define the concept of *satisfy* that is used throughout this thesis in order to determine objects that meet the requirements of a given range query. Fundamentally, given the nature of objects with continuous ranges that will definitely have a probability density function induced over the ranges, we say that an object *satisfies* a range query if its probability to *satisfy* the range query is at least a probability threshold value $\tau$. We define three major possible cases that would most likely occur between a pair of objects with *uncertain dimensions* that intersect with the range query in order to compute the probability of a dominating object. Based on the above concept and possible cases, we define the skyline on *uncertain dimensions* with respect to the range query.

We develop an extension to the *SkyQUD* framework, named *SkyQUD-T*, in order to be able to accommodate skyline with range queries on *uncertain dimensions*. Following the aforementioned concept, a method, *Range Pruning*, which will filter objects that $\tau$-*satisfy* a given range query has been developed. The *Range Pruning* method is designed to filter out objects that do not $\tau$-*satisfy* a given range that is queried on an *uncertain dimension* before any skyline computations is performed. When a set of objects that $\tau$-*satisfy* the range query is returned by the *Range Pruning* method, the *SkyQUD-T* framework will then determine the skyline on these objects. The remaining methods in the *SkyQUD-T* framework follow the exact execution as the methods in the *SkyQUD* framework. The only difference being when computing the probabilities of objects to not be dominated by other objects, we have to take into consideration the range query as well.

## 1.7 Thesis Organisation

The remainder of the thesis is organized as follows:

**Chapter 1** is an introductory chapter that discusses the motivations behind this research study, the challenges and contributions of the research, the objectives, and the scope of the research.

**Chapter 2** reviews the background and the literature review relevant to the research study. The background presents the fundamental concepts of the preference queries and the preference evaluation techniques in the database. It reviews relevant works proposed by previous researchers of preference queries including Top-*k*, Skyline, Top-*k* dominating, and *k*-dominance. The concept of skyline processing techniques is given focus and the techniques are classified based on data certainty, namely: certain and uncertain data. The chapter illuminates as well on the fundamental concept of probabilistic databases and probability theory that is relevant to this research study. The chapter reviews relevant works proposed by previous researchers on existing techniques of preference queries in database environment. It presents the features of these preference techniques, the strengths, and the weaknesses of them.

**Chapter 3** defines how this research study was conducted. This chapter discusses the different phases in this research work and the methodology followed during each phase. The performance metrics, the settings of the experiments, and the datasets that have been used in the experiments of this research study are presented as well.

**Chapter 4** presents in detail the depiction of the proposed framework for processing skyline queries on uncertain data. This chapter also explains and discusses the phases of the proposed framework with a running example of database.

**Chapter 5** elucidates in depth the steps of the proposed extension to the previous framework for processing skyline with range queries on uncertain data. This chapter illustrates the new method that is added as an extension and the existing methods that are modified to accommodate range queries, which are presented and clarified using example database.

**Chapter 6** presents the results of the systematic performance study conducted to evaluate the performance of the proposed frameworks compared to the most relevant existing algorithms. The chapter also discusses the results with respect to different parameters including, the size of the dataset, the threshold value, the distributions of exact values and continuous ranges in a dimension, and the number of dimensions.

**Chapter 7** reflects the conclusions and the contributions of this research study. In addition, the recommendations of the future works are presented in this chapter.

# REFERENCES


Abiteboul, S., Kanellakis, P., and Grahne, G., 1987. On the Representation and Querying of Sets of Possible Worlds, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 34-48.

Akbarinia, R., Pacitti, E., and Valduriez, P., 2007. Processing Top-*k* Queries in Distributed Hash Tables, *Proceedings of the 13th International Euro-Par Conference*, pp. 489-502.

Alwan, A. A., Ibrahim, H., and Udzir, N. I., 2014. A Framework for Identifying Skylines over Incomplete Data, *3rd International Conference on Advanced Computer Science Applications and Technologies*, pp. 79-84.

Atallah, M. and Qi, Y., 2009. Computing All Skyline Probabilities for Uncertain Data, *Proceedings of the ACM SIGMOD-SIGACT-SIGART Symposium of the Principles of Database Systems* (PODS), pp. 279-287.

Babcock, B. and Olston, C., 2003. Distributed Top-*k* Monitoring, *Proceedings of the International Conference on Management of Data*, pp. 28-39.

Balke, W., G¨untzer, U., and Zheng, J. X., 2004. Efficient Distributed Skylining for Web Information Systems, *Proceedings of International Conference on Extending Database Technology* (EDBT), pp. 256–273.

Bartolini, I., Ciaccia, P., and Patella, M., 2006. SaLSa: Computing the Skyline without Scanning the Whole Sky, *Proceedings of the 15th International Conference on Information and Knowledge Management (ICIKM06)*, pp. 405- 414.

Bartolini, I., Ciaccia, P., and Patella, M., 2013. The Skyline of a Probabilistic Relation, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 7, pp. 1656- 1669.

Berchtold, S., Keim, D. A., and Kriegel, H. P., 1996. The X-tree: An Index Structure for High-Dimensional Data, *Proceedings of the 22nd International Conference on Very Large Data Bases* (VLDB), pp. 28-39.

Bhattacharya, A., Teja, P., and Dutta, S., 2011. Caching Stars in the Sky: A Semantic Caching Approach to Accelerate Skyline Queries, *Proceedings of the International Conference on Database and Expert Systems Applications* (DEXA), pp. 493–501.

Böhm, C., Fiedler, F., and Oswald, A., Plant, C., and Wackersreuther, B., 2009. Probabilistic Skyline Queries, *Proceedings of the 18th ACM conference on Information and knowledge management* (CIKM), pp. 651-660.

Börzsönyi S., Kossmann D., and Stocker K., 2001. The Skyline Operator, *Proceedings of the 17th International Conference on Data Engineering* (ICDE), pp. 421-430.

Bruno, N., Chaudhuri S., and Gravano L., 2002a. Top-*k* Selection Queries over Relational Databases: Mapping Strategies and Performance Evaluation, *ACM Transactions on Database Systems*, Vol. 27, No. 2, pp. 153–187.

Bruno, N., Gravano, L., and Marian, A., 2002b. Evaluating Top-*k* Queries over Web-accessible Databases, *Proceedings of the 18th International Conference on Data Engineering*, pp. 369.

Chan, C., Eng, P., and Tan, K., 2005. Stratified Computation of Skylines with Partially-Ordered Domains, *Proceedings of International Conference on Management of Data* (SIGMOD), pp. 203–214.

Chan, C.-Y., Jagadish, H. V., Tan, K.-L., Tung, A. K. H., and Zhang, Z., 2006. Finding k-dominant Skylines in High Dimensional Space, *Proceedings of International Conference on Management of Data* (SIGMOD), pp. 503–514.

Chang, K. C. and Hwang, S., 2002. Minimal Probing: Supporting Expensive Predicates for Top-*k* Queries, *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 346– 357.

Chang, Y., Bergman, L. D., Castelli, V., Li, C., Lo, M., and Smith, J. R., 2000. The Onion Technique: Indexing for Linear Optimization Queries, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pp. 391–402.

Chaudhuri, S. and Gravano, L., 1999. Evaluating Top-*k* Selection Queries, *Proceedings of the 25th International Conference on Very Large Data Bases* (VLDB), pp. 397-410.

Chaudhuri, S., Gravano, L., and Marian, A., 2004. Optimizing Top-*k* Selection Queries over Multimedia Repositories, *IEEE Transaction Knowledge and Data Engineering*, Vol. 16, No. 8, pp. 992-1009.

Chester, S., Mortensen, M. L., and Assent, I., 2014. On the Suitability of Skyline Queries for Data Exploration, *Workshop Proceedings of the International Conference on Extending Database Technology* (EDBT), pp. 161-166.

Chomicki, J., Godfrey, P., Gryz, J., and Liang, D., 2003. Skyline with Presorting, *Proceedings of International Conference on Data Engineering* (ICDE), pp. 717–816.

Cui, B., Lu, H., Xu, Q., Chen, L., Dai, Y., and Zhou, Y., 2008. Parallel Distributed Processing of Constrained Skyline Queries by Filtering, *Proceedings of International Conference on Data Engineering* (ICDE), pp. 546–555.

Das, G., Gunopulos, D., Koudas, N., and Tsirogiannis, D., 2006. Answering Top-*k* Queries Using Views, *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 451–462.

Dellis, E. and Seeger, B., 2007. Efficient Computation of Reverse Skyline Queries, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 291–302.

Ding, X., Lian, X., Chen, L., and Jin, H., 2012. Continuous Monitoring of Skylines over Uncertain Data Streams, *Information Science*, Vol. 184, pp. 196-214.

Elmi, S., Benouaret, K., Hadjali, A., Tobji, M. A. B., and Yaghlane B. B., 2014. Computing Skyline from Evidential Data, *Scalable Uncertainty Management: 8th International Conference* (SUM), pp. 148-161.

Elmi, S., Tobji, M. A. B., Hadjali, A., and Yaghlane B. B., 2017. Selecting Skyline Stars over Uncertain Databases: Semantics and Refining Methods in the Evidence Theory Setting, *Applied Soft Computing*, Vol. 57, pp. 88-101.

Fagin, R., Lotem, A., and Naor, M., 2001. Optimal Aggregation Algorithms for Middleware, *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (PODS), pp. 102-113.

Fagin, R., Kumar, R., and Sivakumar, D., 2003. Comparing Top *k* Lists, *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 28–36.

Godfrey, P., Shipley, R., and Gryz, J., 2005. Maximal Vector Computation in Large Data Sets, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 229–240.

Grahne, G., 1991. The Problem of Incomplete Information in Relational Databases, *Springer-Verlag*.

Green, T. and Tannen, V., 2006. Models for Incomplete and Probabilistic Information, *Proceedings of the International Conference on Current Trends in Database Technology* (EDBT), pp. 278-296.

Groz, B. and Milo, T., 2015. Skyline Queries with Noisy Comparisons. *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (PODS), pp. 185-198.

Güntzer, U., Balke, W., and Kießling, W., 2000. Optimizing Multi-feature Queries for Image Databases, *Proceedings of the 26th International Conference on Very Large Data Bases* (VLDB), pp. 419–428.

Güntzer, U., Balke, W., and Kießling, W., 2001. Towards Efficient Multi-feature Queries in Heterogeneous Environments, *Proceedings of the*

*International Conference on Information Technology: Coding and Computing*, p. 622.

Gupta, R. and Sarawagi, S., 2006. Creating Probabilistic Databases from Information Extraction Models, *Proceedings of the 32nd International Conference on Very Large Data Bases*, pp. 965–976.

Haghani, P., Michel, S., and Aberer, K., 2009. Evaluating Top-*k* Queries over Incomplete Data Streams, *Proceedings of the 18th International Conference on Information and Knowledge Management* (CIKM), pp. 877-886.

Han, X., Li, J., Yang, D., and Wang, J., 2013. Efficient Skyline Computation on Big Data, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 25, No. 11, pp. 2521-2535.

He, G., Chen, L., Zeng, C., Zheng, Q., and Zhou, G., 2016. Probabilistic Skyline Queries on Uncertain Time Series, *Neurocomputing*, Vol. 191, pp. 224-237.

Hristidis, V., Koudas, N., and Papakonstantinou, Y., 2001. PREFER: A System for the Efficient Execution of Multi-parametric Ranked Queries, *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, pp. 259–270.

Huang, Y.-K., 2017. Continuous $d_\varepsilon$-Skyline Queries for Objects with Time-Varying Attribute in Road Networks, *Proceedings of the IEEE 31st International Conference on Advanced Information Networking and Applications*, pp. 439-446.

Ilyas, I. F., Aref, W. G., and Elmagarmid, A. K., 2002. Joining Ranked Inputs in Practice, *Proceedings of the 28th International Conference on Very Large Data Bases*, pp. 950–961.

Ilyas, F. I., Walid, G. A., and Elmagarmid, A. K., 2003. Supporting Top-*k* Join Queries in Relational Databases, *Proceedings of the 29th International Conference on Very Large Databases*, pp. 754–765.

Imielinski, T. and Lipski, W. Jr., 1984. Incomplete Information in Relational Databases, *Journal of the ACM*, Vol. 31, No. 4, pp. 761-791.

Jampani, R., Xu, F., Wu, M., Perez, L. L., Jermaine, C., and Haas, P. J., 2008. MCDB: A Monte Carlo Approach to Managing Uncertain Data, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 687–700.

Jiang, B. and Pei, J., 2009. Online Interval Skyline Queries on Time Series, *Proceedings of the 25th International Conference on Data Engineering* (ICDE), pp. 1036–1047.

Khalefa, M. E., Mokbel, M. F., and Levandoski, J. J., 2008. Skyline Query Processing for Incomplete Data, *Proceedings of International Conference on Data Engineering* (ICDE), pp. 556–565.

Khalefa, M. E., Mokbel, M. F., and Levandoski, J. J., 2010. Skyline Query Processing for Uncertain Data, *Proceedings of the 19th ACM International Conference on Information and Knowledge Management* (CIKM), pp. 1293-1296.

Kießling, W., 2002. Foundations of Preferences in Database Systems, *Proceedings of the International Conference on Very Large Data Bases* (VLDB), pp. 311-322.

Kim, D., Im, H., and Park, S., 2012. Computing Exact Skyline Probabilities for Uncertain Databases, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 12, pp. 2113-2126.

Kontaki, M., Papadopoulos, A. N., and Manolopoulos, Y., 2008. Continuous *k*-dominant Skyline Computation on Multidimensional Data Streams, *Proceedings of the 2008 ACM Symposium on Applied Computing* (SAC), pp. 956-960.

Kontaki, M., Papadopoulos, A. N., and Manolopoulos. Y., 2010. Continuous Processing of Preference Queries in Data Streams, *Proceedings of the 36th International Conference on Current Trends in Theory and Practice of Computer Science*, pp. 47-60.

Kontaki, M., Papadopoulos, A. N., and Manolopoulos, Y., 2012. Continuous Top-*k* Dominating Queries, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 5, pp. 840-853.

Kossmann, D., Ramsak, F., and Rost, S., 2002. Shooting Stars in the Sky: An Online Algorithm for Skyline Queries, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 275–286.

Koutrika, G. and Ioannidis, Y. E., 2004. Personalization of Queries in Database Systems, *Proceedings of the International Conference on Data Engineering*, pp. 597-608.

Kulkarni, R. D. and Momin, B. F., 2016. Parallel Skyline Computation for Frequent Queries in Distributed Environment, *International Conference on Computational Techniques in Information and Communication Technologies* (ICCTICT), pp. 374-380.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N., 2011. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, *Proceedings of the 18th International Conference on Machine Learning*, pp. 282–289.

Le, T. M. N., Cao, J., and He, Z., 2016. Answering Skyline Queries on Probabilistic data Using the Dominance of Probabilistic Skyline Tuples, *Information Science*, Vol. 340, pp. 58-85.

Lee, K. C. K., Lee, W.-C., Zheng, B., Li, H., and Tian, Y., 2010. Z-SKY: An Efficient Skyline Query Processing Framework Based on Z-order. *Journal of Very Large Database*, *VLDB*, Vol. 19, No. 3, pp. 333-362.

Li, C., Chang, K. C., Ilyas, I. F., and Song, S., 2005. RankSQL: Query Algebra and Optimization for Relational Top-*k* Queries, *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pp. 131–142.

Li, C., Chang, K. C., and Ilyas, I. F., 2006. Supporting Ad-hoc Ranking Aggregates, *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pp. 61–72.

Li, X., Wang, Y., Li, X., and Wang, G., 2012. Skyline Query Processing on Interval Uncertain Data, *IEEE 15th International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing Workshops*, pp. 87-92.

Li, X., Ren, K., Li, X., and Yu, J., 2017. Efficient Skyline Computation over Distributed Interval Data, *Concurrency and Computation: Practice and Experience*, Vol. 29, No. 10, pp. 1-19.

Lin, X., Yuan, Y., Wang, W., and Lu, H., 2005. Stabbing the Sky: Efficient Skyline Computation over Sliding Windows, *Proceedings of International Conference on Data Engineering* (ICDE), pp. 502–513.

Lin, X., Xu, J., and Hu, H., 2013. Range-based Skyline Queries in Mobile Environments, *IEEE Transaction on Knowledge and Data Engineering*, Vol. 25, No. 4, pp. 835-849.

Lo, E., Yip, K.Y., Lin, K.I., and Cheung, D.W., 2006. Progressive Skylining over Web-accessible Databases, *Data Knowledge Engineering* (DKE) 57(2), pp. 122–147.

Matteis, T. D., Girolamo, S. D., and Mencagli, G., 2016. Continuous Skyline Queries on Multicore Architectures, *Concurrency and Computation: Practice and Experience*, Vol. 28, No. 12, pp. 3503-3522.

Michel, S., Triantafillou, P., and Weikum, G., 2005. KLEE: A Framework for Distributed Top-*k* Query Algorithms, *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 637–648.

Mortensen, M. L., Chester, S., Assent, I., and Magnani, M., 2015. Efficient Caching for Constrained Skyline Queries, *Proceedings of the International Conference on Extending Database Technology* (EDBT), pp. 1-12.

Mouratidis, K., Bakiras, S., and Papadias, D., 2006. Continuous Monitoring of Top-*k* Queries over Sliding Windows, *Proceedings of the International Conference on Management of Data* (ICMD), pp. 635-646.

Natsev, A., Chang, Y., Smith, J. R., Li, C., and Vitter, J. S., 2001. Supporting Incremental Join Queries on Ranked Inputs, *Proceedings of the 27th International Conference on Very Large Data Bases* (VLDB), pp. 281–290.

Papadias, D., Tao, Y., Fu, G., and Seeger, B., 2003. An Optimal and Progressive Algorithm for Skyline Queries, *Proceedings of International Conference on Management of Data* (SIGMOD), pp. 467–478.

Papadias, D., Tao, Y., Fu, G., and Seeger, B., 2005. Progressive Skyline Computation in Database Systems, *ACM Transactions on Database Systems*, Vol. 30, No. 1, pp. 41–82.

Pei, J., Jin, W., Ester, M., and Tao, Y., 2005. Catching the Best Views of Skyline: A Semantic Approach Based on Decisive Subspaces, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 253–264.

Pei, J., Jiang, B., Lin, X., and Yuan, Y., 2007. Probabilistic Skylines on Uncertain Data, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 15–26.

Rahul, S. and Janardan, R., 2012. Algorithms for Range-skyline Queries, *Proceedings of the 20th International Conference on Advances in Geographic Information Systems* (SIGSPATIAL), pp. 526-529.

Ré, C., Dalvi, N. N., and Suciu, D., 2007. Efficient Top-*k* Query Evaluation on Probabilistic Data, *Proceedings of the 23rd International Conference on Data Engineering*, pp. 886–895.

Ré, C., Letchner, J., Balazinksa, M., and Suciu, D., 2008. Event Queries on Correlated Probabilistic Streams. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 715–728,
2008.

Ross, S. M., 2003. Introduction to Probability Models, Eighth Edition, *American Press*.

Sadri, F., 1991. Modeling Uncertainty in Databases, *Proceedings of the IEEE 7th International Conference on Data Engineering* (ICDE), pp. 122-131.

Sheng, C. and Tao, Y., 2012. Worst-Case I/O-Efficient Skyline Algorithms, *ACM Transactions on Database Systems*, Vol. 37, No. 4, pp. 1-24.

Siddique, M. A., and Morimoto, Y., 2009. *k*-dominant Skyline Computation by Using Sort-Filtering Method, *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 839-848.

Soliman, M. A., Ilyas, I. F., and Chang, K. C.-C., 2007. Top-*k* Query Processing in Uncertain Databases, *Proceedings of the 23rd International Conference on Data Engineering*, pp. 896–905.

Soliman, M. A., Ilyas, I. F., and Ben-David, S., 2010. Supporting Ranking Queries on Uncertain and Incomplete data, *The Very Large Database Journal, VLDB*, Vol. 19, No. 4, pp. 477- 501.

Su, H. Z., Wang, E. T., Chen, A. L. P., 2010. Continuous Probabilistic Skyline Queries over Uncertain Data Streams, *Proceedings of the International Conference on Database and Expert Systems Applications* (DEXA), pp. 105–121.

Suciu, D., Olteanu, D., Ré, C., and Koch, C., 2011. Probabilistic Databases, *Morgan & Claypool*.

Sun, S., Huang, Z., Zhong, H., Dai, D., Liu, H., and Li, J., 2010. Efficient Monitoring of Skyline Queries over Distributed Data Streams, *Knowledge and Information Systems* 25, 575–606.

Tan, K. L., Eng, P. K., and Ooi, B. C., 2001. Efficient Progressive Skyline Computation, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 301–310.

Tao, Y., Xiao, X., and Pei, J., 2006. SUBSKY: Efficient Computation of Skylines in Subspaces. *Proceedings of the 22nd International Conference on Data Engineering, (ICDE 2006)*, pp. 65-74.

Theobald, M., Schenkel, R., and Weikum, G., 2005. An Efficient and Versatile Query Engine for TopX Search, *Proceedings of the 31st International Conference on Very Large Data Bases*, pp. 625–636.

Valkanas, G. and Papadopoulos, A., 2010. Efficient and Adaptive Distributed Skyline Computation, *International Conference on Scientific and Statistical Database Management* (SSDBM), pp. 24–41.

Vlachou, A., Doulkeridis, C., and Kotidis, Y., 2008a. Angle-based Space Partitioning for Efficient Parallel Skyline Computation, *Proceedings of International Conference on Management of Data* (SIGMOD), pp. 227–238.

Vlachou, A., Doulkeridis, C., Nørvåg, K., and Vazirgiannis, M., 2008b. On Efficient Top-*k* Query Processing in Highly Distributed Environments, *Proceedings of the International Conference on Management of Data* (ICMD), pp. 753-764.

Vlachou, A. and Nørvåg, K., 2009. Bandwidth-constrained Distributed Skyline Computation, *Proceedings of the International Workshop on Data Engineering for Wireless and Mobile Access* (MobiDE), pp. 17–24.

Wang, W.-C., Wang, E. T., and Chen, A. L. P., 2011. Dynamic Skylines Considering Range Queries, *Proceedings of the 16th International Conference on Database Systems for Advanced Applications*, pp. 235–250.

Wolfson, O., Sistla, A. P., Chamberlain, S., and Yesha, Y., 1999. Updating and Querying Databases that Track Mobile Units, *Distributed and Parallel Databases*, Vol. 7, No. 3, pp. 257–387.

Wu, P., Zhang, C., Feng, Y., Zhao, B., Agrawal, D., and Abbadi, A., 2006. Parallelizing Skyline Queries for Scalable Distribution, *Proceedings of International Conference on Extending Database Technology* (EDBT), pp. 112–130.

Yi, K., Yu, H., Yang, J., Xia, G., and Chen. Y., 2003. Efficient Maintenance of Materialized Top-*k* Views, *Proceedings of the 19th International Conference on Data Engineering* (ICDE), pp. 189-200.

Yiu, M. L. and Mamoulis, N., 2007. Efficient Processing of Top-*k* Dominating Queries on Multi-dimensional Data, *Proceedings of the 33rd International Conference on Very Large Data Bases* (VLDB), pp. 483-494.

Yiu, M. L. and Mamoulis, N., 2009. Multi-dimensional Top-*k* Dominating Queries, *The Very Large Database Journal VLDB*, Vol. 18, No. 3, pp. 695-718.

Yuan, Y., Lin, X., Liu, Q., Wang, W., Yu, J.X., and Zhang, Q., 2005. Efficient Computation of the Skyline Cube, *Proceedings of International Conference on Very Large Data Bases* (VLDB), pp. 241–252.

Yong, H., Kim, J.-H., and Hwang. S.-W., 2008. Skyline Ranking for Uncertain Data with Maybe Confidence, *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering Workshop* (ICDEW), pp. 572-579.

Zhang, W., Lin, X., Zhang, Y., Wang, W., Yu, J. X., 2009. Probabilistic Skyline Operator over Sliding Windows, *Proceedings of the International Conference on Data Engineering*, pp. 1060-1071.

Zhang, W., Lin, X., Zhang, Y., Pei, J., and Wang, W., 2010. Threshold-based Probabilistic Top-*k* Dominating Queries, *The VLDB Journal*, Vol. 19, No. 2, pp. 283-305.

Zhang, N., Li, C., Hassan, N., Rajasekaran, S., and Das, G., 2014. On Skyline Groups, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 4, pp. 942-956.

Zhang, Z., Hwang, S., Chang, K. C., Wang, M., Lang, C. A., and Chang, Y., 2006. Boolean + Ranking: Querying a Database by *k*-constrained Optimization, *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pp. 359–370.

Zhao, K., Tao, Y., and Zhou. S., 2007. Efficient Top-*k* Processing in Large-scaled Distributed Environments, *Data and Knowledge Engineering Journal* Vol. 63, No. 2, pp. 315-335.

Zheng, J., Chen, J., and Wang, H., 2017. Efficient Geometric Pruning Strategies for Continuous Skyline Queries, *ISPRS International Journal of Geo-Information*, Vol. 6, No. 3, pp. 1-28.

Zheng, W., Zou, L., Lian, X., Hong, L., and Zhao, D., 2014. Efficient Subgraph Skyline Search over Large Graphs, *ACM International Conference on Information and Knowledge Management*, pp. 1529–1538.

Zhou, X., Li, K., Zhou, Y., and Li, K., 2016. Adaptive Processing for Distributed Skyline Queries over Uncertain Data, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 28, No. 2, pp. 371-384.