Understanding Foundations of Probabilistic Data

Boyang Tang
University of Auckland
Auckland, New Zealand
btan766@aucklanduni.ac.nz

***Abstract*—Over the recent past decade, the volumes of data handled by computers have largely increased. The demands for applications that can handle both deterministic and probabilistic data have also increased due to the changing models of organizational operation and the data needed to facilitate their operations. Specifically, the desires to represent databases whose properties cannot be classified to be deterministic have called for designing and implementation of new and effective models. Besides, uncertain data has been found to be existence in the real world. Probabilistic databases involve using different models to predict a wide array of possible outcomes with each outcomes having a different degree of certainty or uncertainty of its occurrence. This research will explore some of the foundations of probabilistic data by reviewing different scholarly works and deriving a critically analyzed evaluation from them. Moreover, the research will propose frameworks that are essential to facilitate the use of probabilistic data approaches by organizations and show the strengths and weaknesses of the approaches.**

I. Introduction

A majority of databases in the contemporary world hold data with uncertain correctness [1]. Cleaning the uncertain data by cleaning the uncertain data can lead to loss of useful information since a fraction of the uncertain data could be correct [4]. Then again, assuming that all uncertain data is correct could give unseemly results to queries executed on the databases [3]. This means that there is high need of quantifying data stored and processed in such database applications. Though researchers have designed many models to address the issue of uncertain data [4], probabilistic databases have been the most successful in enhancing quantifying the integrity of the data in such databases [5]. Probabilistic databases achieve their functionalities by associating the probability with data items, normally perceived as the measure of sureness that the data is precise [1].

To clearly understand the basis or principles for probabilistic databases, this research will provide a basic overview of the functionalities and the approach that was used to make probabilistic databases a reality. The research will use a descriptive approach where much of the information revealed will be from secondary sources. The study by numerous individuals will facilitate this research to evaluate the strengths and weaknesses of probabilistic databases as well as the differences from other types of database systems. The research will reveal the foundations of probabilistic databases based on the basic ideas and how they operate differently or similarly from the earlier database.

II. Overview of probabilistic database

Probabilistic model in probabilistic data is based on the challenges or rather the weaknesses that were highlighted by numerous researchers on the traditional data approaches. The traditional applications used database systems that required certain or precise semantic data to operate effectively. However, after numerous generations, data management has undergone many paradigm shifts, the biggest obstacle when we processing a large volume of data is the database themselves are incomplete or uncertain, where the need of managing imprecise data or uncertain data has become one of the major shift. There are many areas that have shown that several experiments have challenged the traditional notion that uncertain data was a vague idea and could not exist in the real

world. A probabilistic model was proposed to address the uncertainties in the future, and the idea is designed and supported by numerous theoretical and practical models and algorithms that aim at using earlier and present activities to predict the possible uncertainties that could exist in databases. The state in which probabilistic database could be not deterministic in nature, and therefore there are numerous states in which such databases can exist. Below are some of the key ideas and their definitions that are related to probabilistic data issues.

III. Background And Definitions

This research is based on the foundations of probabilistic data, however, to reveal a clear picture of the principles and models comprised, the research intertwines different arguments and ideas seen from earlier studies.

Probabilities: In all probabilistic database systems each data item is associated with a probability within an interval between (0, 1] where zero is taken to represent the data that is certainly incorrect and therefore not included in the database, whereas 1 represents the data that is correct. The integrity in most probabilistic databases is not represented or associated with data items in an absolute sense but rather taken as the relative measure. For instance, supposing we have associated data items A and B with probabilities P_A and P_B as 0.90 and 0.83 respectively, then it would be incorrect to argue that sources used to obtain data items A are 90% correct and B are 83% correct respectively. The only valid claim for such a case is that database item A have higher chances of being correct when placed in the same context with data items B.

Possible worlds: They are all instances that can be achieved by a probabilistic database. For instance, in a case where we are certain about several tuples and uncertain about other tuples in a database, the state of databases can be designed to contain or not contain some tuples, therefore, giving out numerous possibilities of having some tuples and opting out on others. Breakings down the tuple to different probabilistic databases give numerous tables of possible worlds, which are later combined by taking the sum of all probabilities to one table. However, in practice we cannot enumerate all possible worlds, since there are usually too many, the ***Disjoint-Independent Database*** makes it possible to represent a probabilistic database more simply. This structure of database focuses on how to enumerate possible tuples rather than possible worlds, tuples with distinct keys are independent, and its size totally depends on the size of tuples. For instance, given a probabilistic database PDB with a relation schema R (A, B, C, D), A and B are key attributes, there are serval possible tuples and with different probabilities, and then we group them by their keys [2].

Categories of uncertainty: attribute-level and tuple-level are the two common uncertainties in probabilistic database systems. In tuple level, the chances of a database being at a particular state are represented by the probabilities associated with the worlds. Also, we are not sure of the correctness of the tuple and whether it should be included in the database or not. On the other hand, attribute-level, we are uncertain about the exact values that can be accepted by a tuple or whether the tuple can accommodate one of the numerous possible values.

IV. Related Work

This section describes the earlier researches and studies that have been conducted to expose different ideas on the foundations, challenges, and usefulness of probabilistic data in the contemporary world. The numerous articles show the probabilistic model or rather frameworks and algorithms that are used to facilitate query effectiveness in databases.

A 2011 study by Murthy, Ikeda, and Widom aimed at exploring whether probabilistic databases or uncertain database systems could work along with aggregation [5]. The authors used *Trio* system [5] for probabilistic and uncertain data to describe how aggregation is handled. The study was based on a major challenge seen in uncertain data, where the databases are capable of producing exponentially sized results. The study evaluated the systems by using three alternatives, *low bound*, *high bound*, and *expected value* [5] to experiment their

ability to compute both grouped aggregation queries and full-table queries. The approach helped them to evaluate the scalability and performance of their algorithms that were placed on a large synthetic data set. They used 20 different aggregate functions and implemented them to the Trio system where each function was supported by a full table. By the end of the experiment, the authors revealed that aggregate functions for uncertain data were computed efficiently and were more practical from usability perspectives. Besides, they showed *high* and *low* aggregates as essential indicators of uncertain data. Importantly, based on the challenges or rather the errors, the study suggested that there was need for more research that would reduce complexity seen when handling aggregation queries that involved Trio tables with lineage [5].

A different study by two researchers, Mukherjee and Neogy, aimed at evaluating a probabilistic approach as essentially able to facilitate efficient storage and retrieval of trusted information [4]. The authors used numerous arguments to support their proposal of adopting a data model for management of information that is a swap over during numerous trusted negotiations. The primary motivation of conducting the study was that in the contemporary day there are high expectations that network applications to compare what is transferred between nodes to what the server and client environment expected [4]. Since such activities would require trust, the authors evaluated why a probabilistic approach was the best to build the trust [4]. Their suggestions were supported by numerous algorithms that could be used to estimate and calculate trust within a network application and try to rate worthiness of different nodes using the interactions seen from series of probabilistic ratings using different aspects. They used different scenarios to test the algorithms, for instance, they evaluated an event where server A was used to store the net trust of a networking event at any specific time. The outcomes were used to calculate the probability value by using numerous approaches, for instance, *optimal score approach* [4] and *binomial method* [4] determine the probability of any event based on the various past events and appropriate weight applied to each value in order to minimize the likelihood of a posterior log.

A study conducted by Olteanu and Wen aimed at investigating the "problem of ranking query answers in probabilistic databases [3]." The researchers presumed that many applications of probabilistic databases, users were more interested in ranking answers in the order of few possible answers and the probabilities were mere degrees of data uncertainty and were less meaningful to the users. To conduct their investigations, the authors aimed at using observations from two approaches, first, they were to conduct an experiment that would help them compute a precise ranking of Query answers and use them to approximate each answer's probability. Secondly, they aimed at finding out the probability of query answers sharing common factors that could be used to compute all answers at once. Their experiments used a social network database model and designed tables, which were then assigned data, and different probabilistic events were associated with each tuple. They also introduced a mechanism computing most probable answers by ranking query answers. They used MayBMS-based probabilistic database system engine, SPROUT [3], to implement their computations. The study concluded that the available ranking criterion was expensive because they only ranked 20% to 30% of answers that were based on the number of decomposition steps and execution time [3]. However, introducing probabilistic approach showed a greater shift on the results as all answers were evaluated within the execution time and therefore giving accurate outcome based on the critically separated data. The experiments proved that probabilistic approach was more effective in ranking query answers in a less complex manner [3].

Quantifying the uncertainties as probabilities can provide a clear view of the management of probabilistic data as well as the foundations and challenges. A study by Dalvi and Suciu [2] evaluated different probabilistic database models as well as some theoretical principles to reveal descriptive arguments on existing problems and challenges of managing probabilistic databases. They stated that the most data produced by non-traditional means are uncertain which results the data integration and mappings are difficult.

Basically, data uncertainties can be assumed as two types: it could be that the data itself in real world is imprecise or it could be the sureness of data can be confirmed but people's knowledge about it is unreliable. The primary motive of the study was based on the potential problems that would possibly emerge from the large amounts of data generated by modern systems but were not generated by traditional means. Unlike the previous studies, the authors did not make the assumption that uncertain data exists but rather they argued on paradigms of managing data efficiently irrespective if the real data is deterministic or probabilistic. Based on the studies by AI community [2], the study revealed that inference in probabilistic networks was the primary query related problem and therefore there was a need to provide a distinction for query evaluation and probabilistic inference. After a comparative approach of different models, the authors argued that the degree of heterogeneity and scale of the data were contributing to an increase in the cost of enforcing precise semantics [2]. Through prioritizing uncertainties, it would be possible to curb the high costs and enhance the usability of applications that were previously prohibitive. This study also gave an overlook of challenges and further open problems, such as some special aggregate operations cannot be solved efficiently by a simple ***PDB***, the complexity of computing the probability of this kind of query evaluation is open, and the corresponding study of this area is referred to [5]. Additionally, the problem of ranking should also be worried, when people have incomplete knowledge about the deterministic data, the probabilities are only used for ranking degree of confidence in the data but not meaningful to users, and this open problem was also discussed in [3].

A similar study by Barbara, Garcia-Molina, and Porter evaluated on the effectiveness of managing probabilistic data on slightly different dimension [1]. The researchers developed a probabilistic data model that introduced the probability notion by associating it with missing probabilities. In this model, probabilities associated with the values of attributes, which could be defined as deterministic keys or stochastic values in relations. This means each tuple of a relation must be represented as known real entity. But in most practical situations, some stochastic attributes have no idea about the distribution of their probabilities. This model presents a brilliant feature that is using missing probabilities to account for such incompletely specified probability distribution. For instance, a probability with a wildcard such like 0.4[*], which means 0.4 probability has not been assigned to any particular value in a tuple, just assume this probability is distributed over all ranges, but have no idea about where it is supposed to belong [1]. After giving numerous scenarios and experiments on basic relational operators such as ***Project***, ***Select***, and ***Join***. And then applying their model to derive outcomes, the authors were able to show a probabilistic data model with missing probabilities did give us a great of flexibilities. After that, the study also shows an issue when they were applying this model with missing probabilities, they found the validation or correctness of the operators on miss probabilities are not stable, since the probability ranges may larger than they should be. "Information loss" occurs when two relations are replaced by their ***Join*** relation if the original relation has missing probabilities. Their model aimed at revealing a deeply and descriptive concepts to reveal the preciseness of existence of uncertain data in the real world. Besides, their model was able to represent fuzzy sets that most early models were unable to represent and their focus on missing probability made the study feasible especially when addressing probabilistic database where attribute values were missing.

V. Comparison Of The Studies

The ideas shown by the above studies were similar in that they all revealed one or more aspects that could help readers to understand the primary foundations or rather principles of probabilistic data. However, the studies share several differences especially on the approach they employ to conduct experiments and derive arguments. For instance, while the study by Barbara, Garcia-Molina, and Porter [1] designed a probabilistic model from scratch and showed their arguments from scratch, the other studies used existing models to derive their arguments. This shows that if the earlier models had some aspects of inaccurate algorithm

applications, then all consequent researches would remain stuck on the same error, therefore, making it hard to quantify the challenges and issues for effective probabilistic data management. On the other hand, though the studies reveal their arguments on the principles of probabilistic data, there are no efforts to relate their findings to the contemporary and future organizational requirements except in the study by Dalvi and Suciu [2].

Altogether, the primary strength seen from all the studies is that they provide empirical approaches to support their proposals. The use of algorithms and abstract scenarios makes it easy for the reader to grasp the principles of probabilistic data [4]. Besides, the studies show the appreciation of other author's works through evaluating numerous related works. The approach of stating the major problem evaluated by each research creates a clear picture to the readers and makes it more interesting to keep reading and formulating questions based on the outcomes. The findings, therefore, act as a good source to suggest for future work on making people understand the foundations of probabilistic data.

VI. Concluision

This research is good because its dimensions are based on a real-world problem analysis and especially one of the most advancing sectors in the modern world. The fact that the database systems in the contemporary world are characterized by increasing volumes of data as well as the demand of systems that can process and manage uncertain data effectively makes the research to be more applicable in the current day. Besides, this research has not only focused on proving that uncertain data exist in the real world but has also related the arguments to the problem on the table, and the approach makes it a better reference to future researchers when identifying the areas to evaluate probabilistic data management.

We have exposed a clear understanding of the foundations of probabilistic data where both historic events and contemporary developments tend to reveal the importance of taking the issues and challenges seriously. The future is highly likely to be characterized by applications that generate large volumes of uncertain data, this might be a challenge to database systems as if they fail to respond to the problems posed by the volumes, then the systems are highly likely to fail. In conclusion, numerous gaps exist for researchers to calm the debate between deterministic and probabilistic database systems creating models that will respond to the demands of uncertain data.

References

[1] D. Barbara, H. Garcia-Molina and D. Porter, "The management of probabilistic data", *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 5, pp. 487-502, 1992.

[2] N. Dalvi and D. Suciu, "Management of probabilistic data", *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, PODS '07*, 2007.

[3] D. Olteanu and H. Wen, "Ranking Query Answers in Probabilistic Databases: Complexity and Efficient Algorithms", *2012 IEEE 28th International Conference on Data Engineering*, 2012.

[4] S. Mukherjee and S. Neogy, "Storage & retrieval of trusted information: A temporal probabilistic database approach", *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*, 2015.

[5] R. Murthy, R. Ikeda, and J. Widom, "Making Aggregation Work in Uncertain and Probabilistic Databases", *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 8, pp. 1261-1273, 2011.