NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **On the Quality and Price of Data**

Speaker: Mr Tang Ruiming

Date/Time: 12 September 2014, Friday, 02:00 PM to 03:30 PM

Venue: SR7, COM1-02-07

Supervisor : Dr Stephane Bressan, Associate Professor, School of Computing

Abstract:

Data consumers, data providers and data market owners participate in data markets. Data providers collect, clean and trade data. In this thesis, we study the quality and price of data. More specifically, we study how to improve data quality through conditioning, and the relationship between quality and price of data.

In order to improve data quality (more specifically, accuracy) by adding constraints or information, we study the conditioning problem. A probabilistic database denotes a set of non-probabilistic databases called possible worlds, each of which has a probability. This is often a compact way to represent uncertain data. In addition, direct observations and general knowledge, in the form of constraints, help in refining the probabilities of the possible worlds, and possibly ruling out some of them. Enforcing such constraints on the set of the possible worlds of a probabilistic database, obtaining a subset of the possible worlds which are valid under the given constraints, and refining the probability of each valid possible world to be the conditional probability of the possible world when the constraints are true, is called conditioning the probabilistic database. The conditioning problem is to find a new probabilistic database that denotes the valid possible worlds, with respect to the constraints, with their new probabilities. We propose a framework for representing conditioned probabilistic (relational and XML) data. Unfortunately, the general conditioning problem involves the simplification of general Boolean expressions and is NP-hard. Specific practical families of constraints are thus identified, for which efficient algorithms to perform conditioning are devised and presented.

Data providers and data consumers expect the price of data to be commensurate with its quality. We study the relationship between quality and price of data. We separate the cases wherein data consumers request data items directly, and those in which data consumers specify the parts of data they are interested in by issuing queries. For pricing data items, we propose a pricing framework in which data consumers can trade data quality for discounted prices. For pricing queries, we propose a pricing framework to define, compute and estimate

the prices of queries.

For pricing data items, we propose a theoretical and practical pricing framework for a data market in which data consumers can trade data quality for discounted prices. In most data markets, prices are prescribed and not negotiable, and give access to the best data quality that the data provider can achieve. Instead, we consider a model in which data quality can be traded for discounted prices: ``what you pay for is what you get''. A data consumer proposes a price for the data that she requests. If the price is less than the price set by the data provider, then she will possibly get a lower-quality version of the requested data. The data market owners negotiate the pricing schemes with the data providers. They implement these schemes for generating lower-quality versions of the requested data. We propose a theoretical and practical pricing framework with algorithms for relational data and XML data respectively.

We study the problem of defining and computing the prices of queries for cases wherein data consumers request for data in forms of queries. A generic query pricing model which is based on minimal provenances, i.e., minimal sets of tuples contributing to the query result (which can be viewed as the quality of the query result) is proposed. A data consumer has to pay for the tuples that her query needs to produce the query result: ``what you pay for is what you get''. If a query needs higher-quality (namely higher-price) tuples, the price of this query should be higher. The proposed model fulfills desirable properties, such as contribution monotonicity, bounded-price and contribution arbitrage-freedom. It is found that computing the exact price of a query in our pricing model is NP-hard, and a baseline algorithm to compute the exact price of a query is presented. Several heuristics are devised, presented and compared. A comprehensive experimental study is conducted to show their effectiveness and efficiency.