Systems Aspects of Probabilistic Data Management

Magdalena Balazinska, Christopher Ré and Dan Suciu
Department of Computer Science and Engineering
University of Washington
Seattle, WA, USA
{magda,chrisre,suciu}@cs.washington.edu

1. INTRODUCTION

There has been a wide interest recently in managing probabilistic data [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]. But in order to follow the rich literature on probabilistic databases one is often required to take a detour into probability theory, correlations, conditionals, Monte Carlo simulations, error bounds, topics that have been studied extensively in several areas of Computer Science and Mathematics. Because of that, it is often difficult to get to the algorithmic and systems level aspects of probabilistic data management. In this tutorial, we will distill these aspects from the, often theory-heavy literature on probabilistic databases. We will start by describing a real application at the University of Washington, using the *RFID Ecosystem*; we will show how probabilities arise naturally, and why we need to cope with them. We will then describe what an implementor needs to know to process SQL queries on probabilistic databases. In the second half of the tutorial, we will discuss more advanced issues, such as event processing over probabilistic streams, and views over probabilistic data.

2. OUTLINE

The tutorial is divided into five parts. All topics covered in the tutorial depend on Part I, which introduces motivating applications and the probabilistic data model. However, Part II through IV are independent of each other. Part V will focus on current challenges of probabilistic data management systems.

Part I. Motivating application and the Probabilistic Data Model (30 minutes)

- An illustration of RFID enabled applications using the RFID Ecosystem at the University of Washington [21]. Causes and effects of noisy RFID readings. Impacts of RFID deployment limitations.
- Basic methods for coping with errors from RFID readings: particle filters, filtering, smoothing.
- Probabilistic data model. Probabilistic tuples, probabilistic attributes, and tradeoffs between the two models; Models using independence, disjointness, or correlation. Possible worlds. Data lineage.

PVLDB '08, August 23-28, 2008, Auckland, New Zealand

Part II. General Query Processing Techniques (35 minutes)

.

- Extensional plans in SQL, intensional plans, the "safety" condition for plans [8]; general purpose probabilistic inference based on the Luby-Karp Monte Carlo algorithm.
- Top-K query answering.
- Skyline queries.
- Aggregate queries with group-by and aggregate operators (expected value semantics); queries a GROUP-BY clause Having Clause; OLAP for Probabilistic databases.

Break (30 minutes)

Part III: Processing Probabilistic Events (25 minutes)

- Motivation for and definition of events, event queries, and their semantics on probabilistic data.
- Online event query processing over Markov Chains.
- Offline event query processing over Markov Chains. Indexing archived Markov Chains.

Part IV: Advanced Representation for Probabilistic Databases (30 minutes)

- Representation Techniques for Discrete Distributions: basic Lineage; Views over probabilistic data; World decomposition.
- Probabilistic database systems: Trio, MystiQ, URanks, Orion, Monte-Carlo DB.

Part V: Discussions and Open Problems (10 minutes)

We will describe a few short and long term challenges in probabilistic data management.

3. INTENDED AUDIENCE

This tutorial is primarily intended for researchers, developers, and PhD students. It targets mostly a database audience, but some parts of this tutorial may be of interest to researchers and practitioners in Ubiquitous Computing, while others to experts in Knowledge

Representation. The emphasis of the tutorial will be on the algorithmic and systems building side of probabilistic databases, and not on theoretical results. The tutorial will be self contained, and will assume only a standard background in database systems and basic probability theory.

4. THE AUTHORS

Magdalena Balazinska is an Assistant Professor at the University of Washington and has a PhD from MIT. Her research is on systems aspects of stream data processing, with a special emphasis on applications that process complex streaming data, including noisy RFID data. Prof. Balazinska is a Microsoft Research New Faculty Fellow (2007), received the Rogel Faculty Support Award (2006), and a Microsoft Research Graduate Fellowship (2003-2005).

Christopher Ré is a senior graduate student at the University of Washington. His research is both on theoretical and systems-level aspects of probabilistic data management. Ré lead the team that built the recently released version of MystiQ, a probabilistic database management system.

Dan Suciu is a Professor at the University of Washington. He received his Ph.D. from the University of Pennsylvania in 1995, then was a principal member of the technical staff at AT&T Labs until he joined the University of Washington in 2000. Professor Suciu is conducting research in data management, with an emphasis on topics that arise from sharing data on the Internet, such as management of semistructured and heterogeneous data, data security, and managing data with uncertainties. He is a co-author of the book Data on the Web: from Relations to Semistructured Data and XML, holds six US patents, received the 2000 ACM SIGMOD Best Paper Award, is a recipient of the NSF Career Award and of an Alfred P. Sloan Fellowship.

5. REFERENCES

[1] P. Andritsos, A. Fuxman, and R. J. Miller. Clean answers over dirty databases. In *ICDE*, 2006.

[2] L. Antova, C. Koch, and D. Olteanu. 10ˆ(10ˆ6) worlds and beyond: Efficient representation and processing of incomplete information. In *ICDE*, 2007.

[3] O. Benjelloun, A. Das Sarma, A. Y. Halevy, M. Theobald, and J. Widom. Databases with uncertainty and lineage. *VLDB J.*, 17(2):243–264, 2008.

[4] O. Benjelloun, A. Das Sarma, C. Hayworth, and J. Widom. An introduction to ULDBs and the Trio system. *IEEE Data Eng. Bull*, 29(1):5–16, 2006.

[5] D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan. Efficient allocation algorithms for olap over imprecise data. In *VLDB*, pages 391–402, 2006.

[6] Douglas Burdick, AnHai Doan, Raghu Ramakrishnan, and Shivakumar Vaithyanathan. Olap over imprecise data with domain constraints. In *VLDB*, pages 39–50, 2007.

[7] R. Cheng, D. Kalashnikov, and S. Prabhakar. Evaluating probabilistic queries over imprecise data. In *SIGMOD*, pages 551–562, 2003.

[8] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In *VLDB*, Toronto, Canada, 2004.

[9] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Model-driven data acquisition in sensor networks. In *VLDB*, pages 588–599, 2004.

[10] A. Deshpande, C. Guestrin, S. Madden, J. M. Hellerstein, and W. Hong. Using probabilistic models for data management in acquisitional environments. In *CIDR*, pages 317–328, 2005.

[11] X. Dong, A. Halevy, and C. Yu. Data integration with uncertainty. In *VLDB*, 2007.

[12] R. Gupta and S. Sarawagi. Creating probabilistic databases from information extraction models. In *VLDB*, pages 965–976, 2006.

[13] Ming Hua, Jian Pei, Wenjie Zhang, and Xuemin Lin. Ranking queries on uncertain data: a probabilistic threshold approach. In *SIGMOD Conference*, pages 673–686, 2008.

[14] E. Hung, L. Getoor, and V.S. Subrahmanian. PXML: A probabilistic semistructured data model and algebra. In *ICDE*, 2003.

[15] Ravi Jampani, Fei Xu, Mingxi Wu, Luis Leopoldo Perex, Chris Jermaine, and Peter J. Haas. Mcdb: A monte carlo approach to managing uncertain data, 2008.

[16] T.S. Jayram, S. Kale, and E. Vee. Efficient aggregation algorithms for probabilistic data. In *SODA*, 2007.

[17] S. Jeffery, M. Garofalakis, and M. Franklin. Adaptive cleaning for RFID data streams. In *VLDB*, pages 163–174, 2006.

[18] Bhargav Kanagal and Amol Deshpande. Online filtering, smoothing and probabilistic modeling of streaming data. In *ICDE*, pages 1160–1169, 2008.

[19] Jian Pei, Bin Jiang, Xuemin Lin, and Yidong Yuan. Probabilistic skylines on uncertain data. In *VLDB*, pages 15–26, 2007.

[20] C. Ré, N. Dalvi, and D. Suciu. Efficient Top-k query evaluation on probabilistic data. In *ICDE*, 2007.

[21] C. Ré, J. Letchner, M. Balazinska, and D. Suciu. Event queries on correlated probabilistic streams. In *SIGMOD*, Vancouver, Canada, 2008.

[22] C. Ré and D. Suciu. Materialized views in probabilistic databases for information exchange and query optimization. In *VLDB*, pages 51–62, 2007.

[23] A. Das Sarma, M. Theobald, and J. Widom. Exploiting lineage for confidence computation in uncertain and probabilistic databases. In *ICDE*, 2008.

[24] Prithviraj Sen and Amol Deshpande. Representing and querying correlated tuples in probabilistic databases. In *ICDE*, 2007.

[25] Mohamed A. Soliman, Ihab F. Ilyas, and Kevin Chen-Chuan Chang. Top-k query processing in uncertain databases. In *ICDE*, pages 896–905, 2007.

[26] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *CIDR*, pages 262–276, 2005.