Computer Science Colloquium

Querying Probabilistic Data

Dan Suciu
University of Washington

Abstract: A major challenge in data management to date is how to manage uncertainty in the data; uncertainty may exist because the data was extracted automatically from text, or was derived from the physical world such as RFID data, or was obtained by integrating several data sets using fuzzy matches, or may be the result of complex stochastic models. This has motivated research on probabilistic databases, where uncertainty is modeled using probabilities, and whose goal is to deliver predictable performance for queries on large probabilistic databases. Probabilistic inference is known to be intractable in general, but once we fix a query and considers only the database as variable input, it becomes a specialized problem, which requires a specialized analysis. I will show that Unions of Conjunctive Queries (also known as non-recursive datalog rules) admit a dichotomy: every query is either provably #P hard, or can be evaluated in PTIME. For practical purposes, the most interesting part of this dichotomy is the PTIME algorithm, which relies on the inclusion/exclusion formula. Interestingly, the algorithm succeeds in evaluating in polynomial time some queries for which the underlying Boolean formula does not admit polynomial size OBDDs or FBDDs, or even (we conjecture) a polynomial size d-DNNF.

Bio:
Dan Suciu is a Professor in Computer Science 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 and 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 two books Data on the Web: from Relations to Semistructured Data and XML, 1999, and Probabilistic Databases, 2011. He holds twelve US patents, received the 2000 ACM SIGMOD Best Paper Award, the 2010 PODS Ten Years Best paper award, and is a recipient of the NSF Career Award and of an Alfred P. Sloan Fellowship. Suciu's PhD students Gerome Miklau and Christopher Re received the ACM SIGMOD Best Dissertation Award in 2006 and 2010 respectively, and Nilesh Dalvi was a runner up in 2008.

Monday, November 14, 2011, 2:00 pm
Mathematics and Science Center: W201

Mathematics and Computer Science
Emory University