Advanced Topics in Databases: Data Provenance and Data Citation

MW 1:30-3

Towne 303

In today’s Big Data-driven science, there is a well-acknowledged need for reproducibility, repeatability, and consistent processing of data. Such capabilities require data provenance, a comprehensive record of the inputs, setting, and processing operations that went into producing a result. The course will cover the theory and practice of data provenance: How it is represented and captured (database, workflow, OS-level, network); Connections to graph databases and query languages (e.g. NoSQL solutions such as REDIS, ProQL); Privacy and security issues; Provenance interoperability, the WWW standard PROV and limitations; Partial provenance; and other current research questions.

Closely related to provenance is the issue of data citation. Citation is an essential part of scientific publishing, is used to gauge the trust placed in published information and, for better or for worse, is an important factor in judging academic reputation. Scientific publishing increasingly involves datasets placed in structured yet evolving databases, and accessed through queries. Although standards have been proposed for citing such datasets, it is not well understood how to automatically generate citations. The course will cover the practice of and computational challenges associated with data citation: Exemplars of citeable datasets; Rule-based citation language; Data archiving; Relationship to query rewriting using views; and other related research questions.

Course Format

The course will be a combination of lectures of introductory material in each of these areas, and research papers presented by students. Students will be expected to write summaries of each paper presented in class, present at least one paper, and do a hands-on project implementing some aspect of data provenance or data citation.

The class is open to Ph.D. students, and Masters students who have taken a database course (CIS550 or equivalent).