CSDS 500 Spring 2022 Colloquium

11:30 AM to 12:30 PM
Tuesday, February 24, 2022
Virtual

Zoom Webinar ID: 924 6942 2087
Passcode: 539603

Efficient Management and Querying Over Incomplete Data

Abstract: In the era of Big Data, real-world application data often exhibit unprecedented features such as variety, volume, and velocity. Low data quality has always been one of the most important issues during the data collection, integration, and analysis. In this talk, I will focus on one fundamental, yet challenging, problem about the data quality, that is, the *data incompleteness*. I will first talk about the background of incomplete data, including real applications that involve missing data, incomplete data classification, and queries over incomplete data.

Then, I will discuss various imputation techniques that manage and analyze incomplete data, and consider a specific problem, *online topic-aware entity resolution over incomplete data streams* (TER-iDS), which online imputes incomplete tuples and detects pairs of topic-related matching entities from incomplete data streams. Due to the efficiency requirements of stream processing and characteristics of incomplete data, it is rather challenging to efficiently perform online ER over incomplete stream data. In order to tackle the TER-iDS problem, we propose an effective imputation strategy, carefully design effective pruning methods, as well as indexes/synopsis, and develop an efficient TER-iDS algorithm via index joins. Extensive experiments have been conducted to evaluate the effectiveness and efficiency of our proposed TER-iDS approach over real data sets.

Xiang Lian
Kent State University

Bio: **Dr. Xiang Lian** received the Bachelor degree from the Department of Computer Science and Technology, Nanjing University, in 2003, and the PhD degree in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology, in 2009. He is now an associate professor at the Department of Computer Science, Kent State University. His research interests include query processing over (1) Incomplete, probabilistic, and inconsistent databases, (2) uncertain and certain graph databases, (3) streaming time series, and (4) spatial-temporal databases. He has published 1 book, 1 book chapter, and more than 91 top-tier conference/journal papers in the area of databases. He served as the session chairs of PVLDB'20, ICDE'17, and SIGIR'21, proceeding co-chairs of SIGMOD'14, SIGMOD'15, APWeb-WAIM'17, and WAIM'16, and program committee (PC) members and/or reviewers in more than 45 conferences and journals.

This is to certify that _____attended this seminar. Certified by _____.
Certificates of attendance and other evidence of CPD activity should be retained by the attendee for auditing purposes.

CASE SCHOOL OF ENGINEERING
CASE WESTERN RESERVE UNIVERSITY