CS 784: Foundations of Data Management

Spring 2022
INTRODUCTION

• undergrad in Athens, Greece
• Ph.D. in University of Washington (the other UW)
• at UW-Madison since 2015!

Research Interests

• parallel query processing
• data pricing
• uncertainty in data management
COURSE LOGISTICS
COURSE FORMAT

• Lectures **Tu+Th** 2:30-3:45 pm

• Office Hours: **Th** 1:30-2:30pm or by appointment

• Webpage: [http://pages.cs.wisc.edu/~paris/cs784-s22/](http://pages.cs.wisc.edu/~paris/cs784-s22/)
COURSE STRUCTURE

The course will have two parts:

1. Query Languages + Complexity
2. Advanced Topics: provenance, privacy, uncertainty, stream processing, graph databases

For some lectures I will post notes on the webpage, for others we will focus on specific papers
PREREQUISITES

It will be helpful if you have good knowledge of:

• Databases, SQL, Relational Algebra
• Algorithms
• Complexity
GRADING

• Class participation: 10%
• Homework (3): 30%
• Paper reviews (4): 20%
• Research project: 40%
HOMEWORK

- Individual assignments
- Submitted through Canvas (use Latex!)
- You can use up to 5 late days for all 3 assignments
PAPER REVIEWS

- Read an assigned paper before the lecture
- Submit a brief review of the paper
- Answer a few questions related to the content of the paper
RESEARCH PROJECT

• In groups of 1 to 3 people
• Independent research on any topic related to the course
• Deliverables:
  – 2/12: email groups + tentative ideas
  – 2/28: project proposal
  – 3/28: milestone
  – Last week: project presentations (10% of grade)
  – 5/8: final report
SAMPLE PROJECTS

• A Lightweight Approach to Approximately Query Big Data
• Efficient Multiway Joins on Heterogeneous Parallel Networks
• Materialized Views In Data Warehousing Environments
• Implementing Datalog on an Asynchronous Distributed Dataflow Framework
What is This Class About?
WHAT IS THIS CLASS ABOUT?

• Data is everywhere!
• Managing data is critical:
  – scientific discoveries
  – online services (social networks, online retailers)
  – decision making
• **Databases** are the core technology
• **In this class:**
  – Foundations of data management
CLASSIC DATABASE THEORY

• Conjunctive Queries (i.e., join queries)
• Query containment/equivalence
• Query complexity
  – how fast can we evaluate a join?
  – how big can the result of a join be?
  – are some join queries easier to compute than others?
Datalog is a declarative language that allows us to express larger classes of queries!
QUERY EVALUATION

• How do we evaluate queries in parallel environments?
  – e.g., Spark

• How do we evaluate queries in streaming environments?
How do we deal with uncertain data?

- probabilistic databases
- query answering over dirty data
- data cleaning / repairs
OTHER TOPICS

• Provenance

• Differential Privacy

• Graph Databases