CS698F Advanced Data Management

Instructor: Medha Atre

Objectives

- Smart techniques for large scale data management
  - Google, Facebook, Amazon survive because of these
- Problems and solutions of non-relational data
  - Graphs, text, media, and mixture of all these, e.g., Wikipedia, social networks.
- Challenges of "scale" in these data
  - Our focus will be on very large graphs, e.g., several million nodes and edges (do not fit entirely in memory).

Why is it important?

- Prototypical solutions do not work in real life
  - Google runs its services on massive scale "clusters" of 1000s of commodity computers connected to each other. It also developed tools like *BigTable*.
  - Yahoo developed *Hadoop*.
  - Facebook developed *Cassandra*.
- If you want to work in an industry, you cannot ignore "scale"! :-)

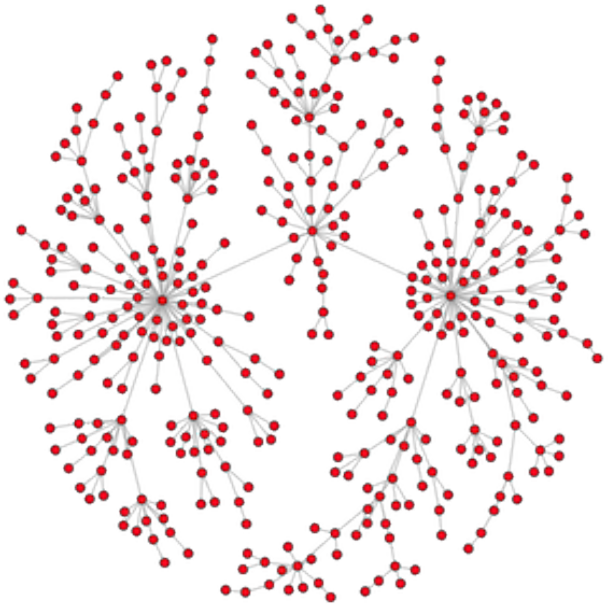
Focus areas

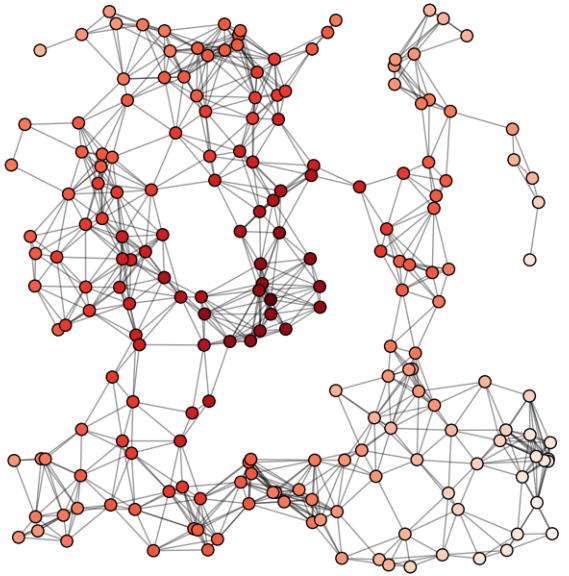
- Some traditional data management, indexing techniques
  - Recap of your UG course on DBMS
- Move to modern techniques
  - Focus on graph data, because that is what internet is made of
- Graph storage techniques – single server and distributed
- Types of graph problems
  - Pattern mining, path queries, keyword search, influence maximization.

Properties of graphs

- *Semi-structured* as opposed to strictly structured data like tabular/relational data
- May have heavy skew
  - Few nodes with lot of edges
  - A lot of nodes with very few edges
- Even structure might be different.. example follows

Structural difference example





Why focus on graphs?

- Entire web is a graph
  - Google's famous *PageRank* algorithm works on internet as a graph.
  - Facebook works by the concept of graphs (social *network*).
  - Twitter is a graph.
  - Amazon product catalogue with users who purchased what is a graph

Why focus on graphs?

- Entire web is a graph
  - Protein-Gene interaction is a graph – bioinformatic networks.
  - IBM's Watson computer in Jeopardy challenge was trained using knowledge-graphs like Yago and DBPedia (RDF graphs).
- Graph based semi-supervised learning (ML techniques)
- In the nutshell – you cannot escape graphs! :-)

Course outline

- Seminar style course management
  - Focus on reading cutting-edge techique papers
  - Individual assignments to develop new techniques or implement existing ones
- Course project
  - Select papers will be floated, and students have freedom to choose topics
  - Topics are expected to be extensions of research papers, their implementation, or brand new ideas (bonus for this)!

Evaluation process

- Will depend on final headcount of the class
  - Grading scheme will be announced closer to add/drop period.
- Attending classes is *strongly* recommended
  - More for your benefit than instructor's :-)
- Slides are supposed to be guidelines and not exhaustive
- Any exams/assignments will be based on what is covered in the classes.

Miscellaneous

- Office hours: by appointment
- My office: KD 219
- While writing to the instructor you **must** add **[CS698F]** at the beginning of the subject line.
  - Helps the instructor to filter out course related emails quickly and respond to the students in a timely manner.
  - Else emails may not get attended and responses might be delayed.