Pick up a handout
on the front table

Welcome to

DS504/CS586: Big Data Analytics --Review

Prof. Yanhua Li

Time: 6:00pm –8:50pm R
Location: AK232
Fall 2016

Today

- 1. Review
  - Key topics, techniques, discussed in the semester
- 2. Future opportunities
  - Big data analytics
  - Urban Computing

10 min Break

- 3. Team 1 presentation
- 4. Course evaluation
- 5. Group discussion for final projects

Introduction

What is "Big Data"?

Big Data Analytics

techniques and tools for managing, analyzing and extracting knowledge from “big data”

CS586/DS504-2016Fall

5. Applications

Urban Computing, Social Network Analysis Networking

4. Big Data Mining

Graph Mining, Data Clustering

Recommender systems, Deep Learning

3. Data Management

Indexing, Query Processing

2. Data Preprocessing/Cleaning

Error Correction, Map-Matching

1. Data Acquisition & Measurement

Representative data collection: Sampling

Techniques

Sampling and index

1. Graph Mining
3. Index, Query
4. Data Collection

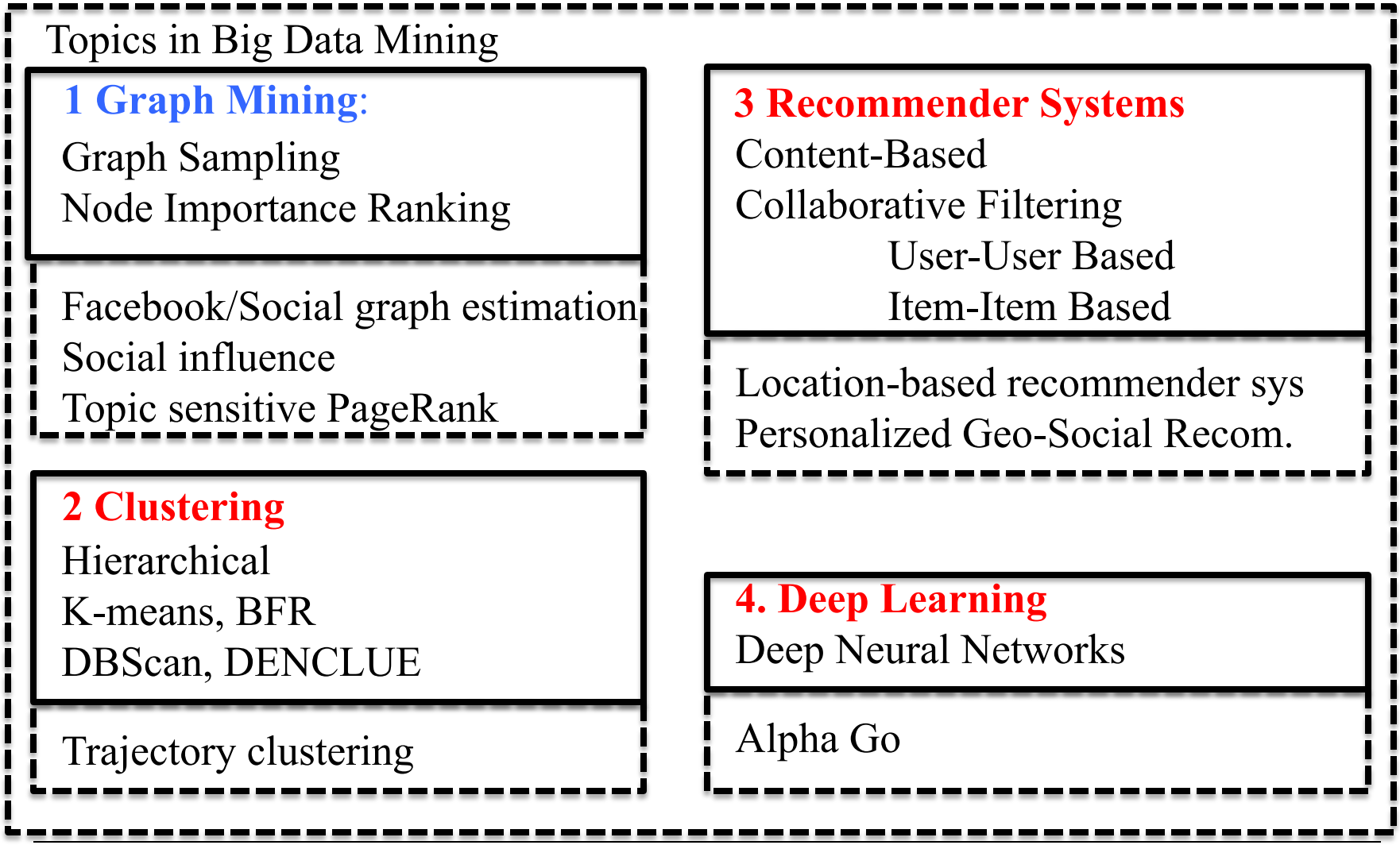
Clustering

4. K-means, DBSCAN
4. BFR, DENCLUE
4. Trajectory Clustering
5. Urban: Bike sharing

More techniques

2. Map-Matching
4. Recommender Systems
4. Deep Learning (Guest)

Big Data Mining Topics



Roadmap

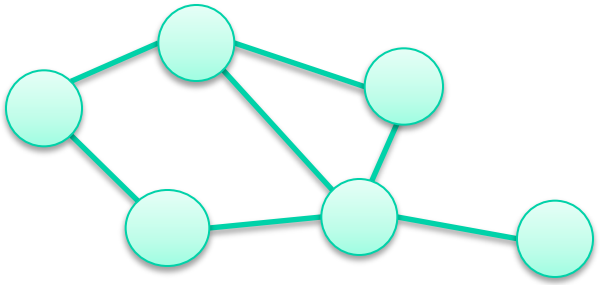
- 1. Sampling & Indexing
  - Random prefix/region/zoomin/region sampling
  - Index structure: B-Tree, Quad-tree, R-tree, etc
- 2. Clustering
  - Hirachical
  - K-means, DBScan
- 3. Recommender System, Deep learning, Map-Matching, etc

- 4. Applications

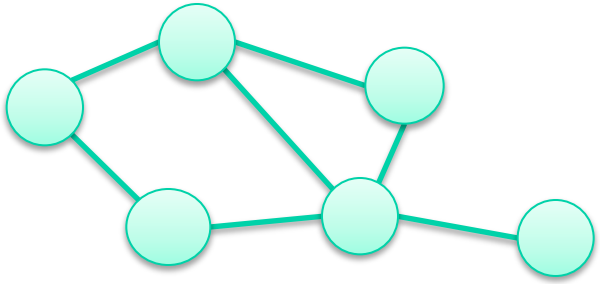
Sampling Big Data

1.1 Random sampling (uniform & independent)

- vertex sampling

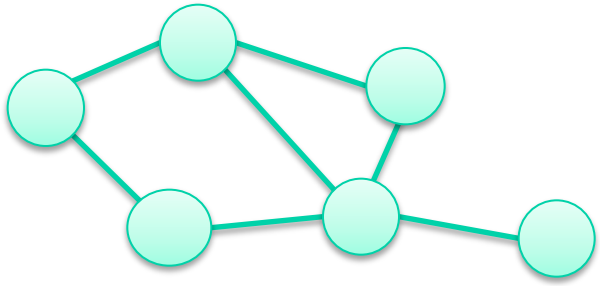


- edge sampling

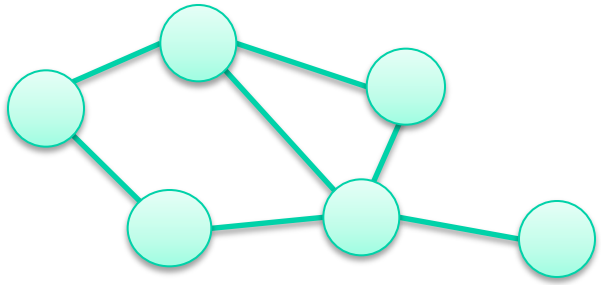


1.2 crawling

- BFS sampling



- random walk sampling



Class Outcomes

What is *DS504/CS586* about?

- We'll learn about
  - Advanced Techniques for Big Data Analytics
    - Large scale data sampling and estimation,
    - Data Cleaning,
    - Graph Data Mining,
    - Data management, clustering, etc.
  - Applications with Big Data Analytics
    - Urban Computing
    - Social network analysis
    - Recommender system, etc.
- Learning outcomes
  - Understand & Explain **challenges** and **advances** in the state-of-art in big data analytics.
  - Design, develop and fully execute a big data analytics **project**.
  - Communicate the **ideas** effectively in the form of a **presentation** and **written documents** to a technical audience.

CS586/DS504-2016Fall

5. Applications

Urban Computing, Social Network Analysis Networking

4. Big Data Mining

Graph Mining, Data Clustering

Recommender systems, Deep Learning

3. Data Management

Indexing, Query Processing

2. Data Preprocessing/Cleaning

Error Correction, Map-Matching

1. Data Acquisition & Measurement

Representative data collection: Sampling

Techniques

Sampling and index

1. Graph Mining
3. Index, Query
4. Data Collection

Clustering

4. K-means, DBSCAN
4. BFR, DENCLUE
4. Trajectory Clustering

More techniques

2. Map-Matching
4. Recommender Systems
4. Deep Learning (Guest)

Project 1 (Single Data Source)

- T1: Allstate Claim Prediction Challenge
- T2: Predicting YouTube 3D Videos Trends
- T3: Sampling Method for Sum Aggregation of Point of Interests on Map
- T4: Mining of Stack Overflow reviews for insights
- T5: Measuring restaurant diversity index for different cities
- T6: GitHub – Sizing up online social networks

Project 2 (Heterogeneous Data)

- T1: Restaurants Location Recommendation
- T2: Online learning performance vs Offline Geographic Information
- T3: Airbnb user behavior prediction
- T4: Community detection in large networks
- T5: Demand-Supply analysis on Regional Restaurant Distribution
- T6: Social Network Marketing through Influence Prediction

- Real application problems
- Data collection/processing/management/mining/ evaluation/visualization/

Workload

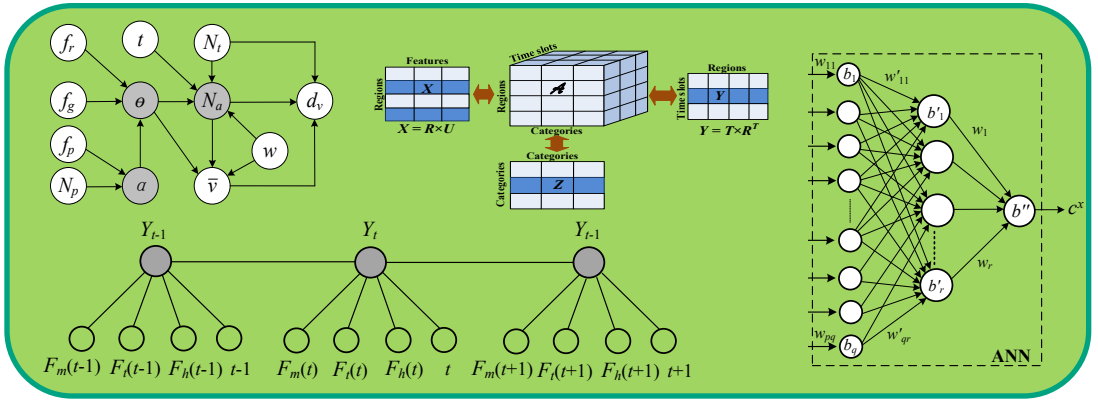
- Focus more on critical thinking, problem solving, “heads-on/hands-on” experiences!
  - Understand, formulate and solve problems
  - Read and critique research papers
  - Two Course Projects
  - Oral presentation
  - Team Work,
  - Coding,

Workload and Grading

- Grading
  - Projects (40%)
    - Project 1 (10%)
    - Project 2 (30%)
      - **Final reports in the discussion forum (by 11:59pm 12/13);**
      - **Self-and-peer evaluation form for project 2 (by 11:59PM 12/13);**
  - Written work (30%):
    - Critiques + Project reports (20%)
    - Quiz (10%, with 5% each)
  - Oral work (30%):
    - Presentation

Data Scientist

Models and Algorithms



Problems

Data

Next Session: Final Project Presentation

- 12/15 R
- 22 min each team (including Q&A)
  - Team 1
  - Team 2
  - Team 3
  - Team 4
  - Team 5
  - Team 6

- Snacks and Drinks will be provided.

Want to learn more?

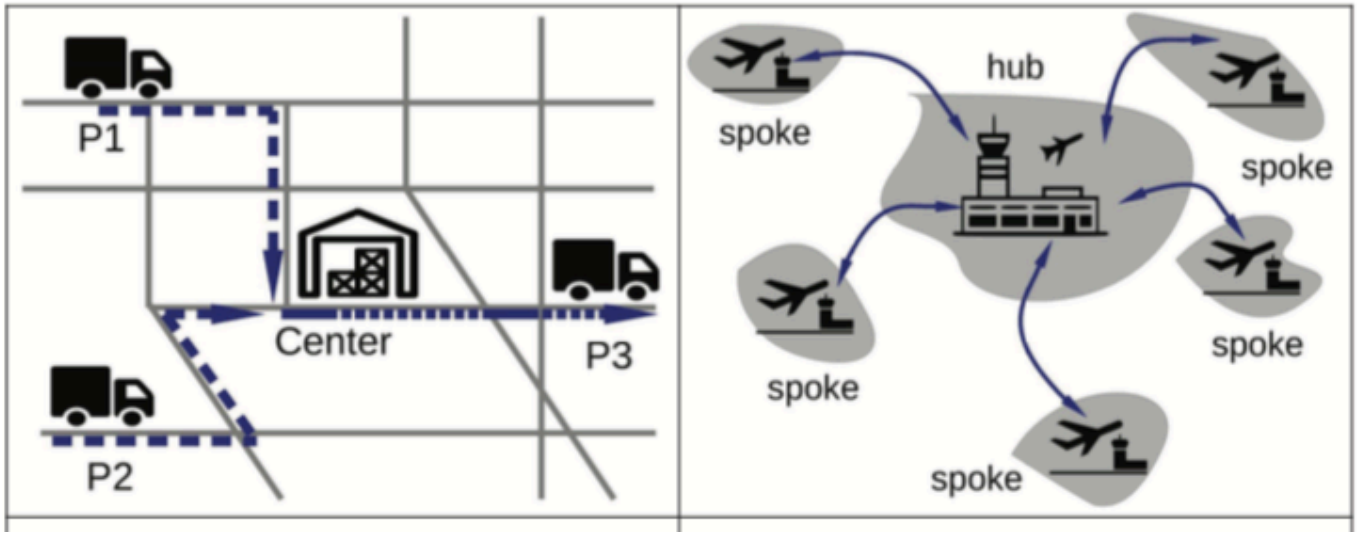
Future Opportunities.

Spring 2017

- DS595/CS525 Special Topics in DS/CS,
  - Urban Computing, applications and methodologies

Urban Computing Research Group at WPI

- Hub-and-Spoke Urban Transportation

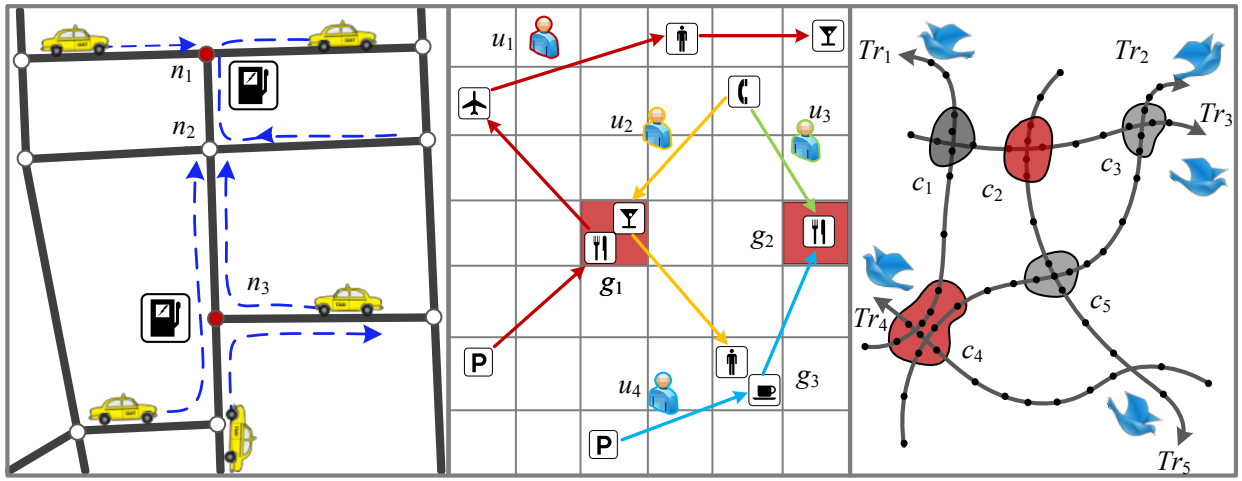


Microsoft Research
微软亚洲研究院

WPI

Urban Computing Research Group at WPI

- Most influential k-location Mining



(a) EV Charging Station Placement

(b) Advertisement Placement

(c) Observation Station Placement

Microsoft Research 微软亚洲研究院

WPI

Urban Computing Research Group at WPI

- Human-in-Loop Urban Computing

Microsoft Research
微软亚洲研究院

WPI

Research opportunities are available in my group.

Contact: yli15@wpi.edu
website: http://wpi.edu/~yli15/index.html