Homework 1:
Time Series Analysis & Network Analysis

Mobile Data Mining

Spring 2019

Goal

- Use Spark, one of the most popular big data processing system to deal with traffic data.
- Use time series model to analyze the traffic data.
- Use network model to analyze the traffic data.

Data

- Mobile network usage traces of Shanghai
  - Involving over 1000 base station (BS)
  - From Aug 1st and Aug 31st 2014

- Format (each line)

Device's ID||Start time||End time|| Location(base station ID)||Traffic volume (Bytes)

- Ethics
  - The dataset describes users' fine-grained behavior, which has privacy implication.
  - Please keep the data on the server, and DO NOT copy any out of the server (the server logs full records of your actions)
  - Please sign the non-disclosure agreement (NDA) in the attachment, and submit a scanned copy with the homework.

Experiments #1: Basic Analysis

- Data Statistics
  - #records, #locations, #user, …
  - Average traffic consumption of each user and each location

- Distribution Analysis
  - User distribution in terms of locations;
  - Traffic consumption distribution in terms of locations;
  - Traffic consumption distribution in terms of users.

Experiments #2: Time Series Analysis

Select the top 3 BSs with the largest traffic

- Time Distribution Statistics
  - Plot the traffic volume of the BSs in different time-bins (1 hour).
- Time Series Decomposition
  - Decompose their traffic into trend component, periodical component, and residual component.
  - Plot the three components of each BS respectively.
- Frequency Analysis **(Bonus: 2 Points)**
  - Implement discrete Fourier transform to their traffic.
  - Plot the amplitude of the obtained Fourier series.
  - Plot the power spectrum of their traffic.

Experiment #3: Network Analysis

- Construct Users' Contact Graph
  - If two users visit the same location within a short time period (1 hour in this experiment), these exists an edge between them;
  - The weight of edge is described by the number of their "encountering".
- Graph Analysis
  - Computing the graph metrics of #nodes, #links, average degree, graph diameter, average path length;
  - Plot the complementary cumulative distribution function (CCDF) of users' degree in the contact graph, and using a suitable distribution to do Curve fitting;
  - Calculate the clustering coefficient of the top 5 users with the largest node degree.
- Community Detection **(Bonus:3 Points)**
  - Using a community detection algorithm introduced in the class, and virtualize the obtained results;
  - Plot the CCDF of the number of users in each community.

Submission

- Submit this homework before March 31st. (Hard Deadline, please keep in mind)
  - **Bonus is not required to be finished.**

- Submit as .zip file, including:
  - 1) A word document, Including:
    - Brief summary about the algorithms designed for data processing
    - All the results you obtained, presented in table or figure (using figure more, and show the results clearly and beautifully)
    - Interpretation/discussion for each result
    - Do not need to copy the code into this document
  - 2) Source code, Including:
    - Spark code
    - Other analysis code
  - 3) A scanned copy of the signed NDA (make sure to have it)

Thank you!