Homework 4

EE232E - Graphs and Network Flows

Due June 9, 2017

Submission: Please submit a zip file containing your codes and report to "ee232e.spring2017@gmail.com". The zip file should be named as "HW1_UID1_UID2_..._UIDn.zip" where UIDx are student ID numbers of team members. If you had any questions you can send an email to the same address.

In this assignment, we will study data from stock market. The data is available on this *Dropbox Link* [1]. The goal of this assignment is to study correlation structures among fluctuation patterns of stock prices using tools from graph theory. The intuition being that investors will have similar strategies of investment for stocks that are effected by the same economic factors. For example, the stocks belonging the transportation sector may have different absolute prices, but if for example fuel prices change or are expected to change significantly in the near future, then you would expect the investors to buy or sell all stocks similarly and maximize their returns. Towards that goal, we construct different graphs based on similarities among the time series of returns on different stocks at different time scales (day vs a week). Then, we study properties of such graphs. The data is obtained from Yahoo Finance website for the last 3 years. You're provided with a number of csv tables, each containing several fields: Date, Open, High, Low, Close,

[1] https://www.dropbox.com/s/83l60htndqpn3fv/finance_data.zip?dl=0

Volume, and Adj Close price. The files are named according to *Ticker Symbol* of each stock. You may find the market sector for each company in `Name_sector.csv`.

A second goal is to utilize the underlying data to study approximation algorithms for solving the Δ-Traveling Salesman Problem (TSP) using Minimum Spanning Trees and Eulerian Cycles.

Determining Return Correlations:

1. **Calculating Correlations among Time Series Data:** The cross correlation coefficient of two different stock-return time series is defined as $\rho_{ij} = \frac{<r_i(t)r_j(t)> - <r_i(t)><r_j(t)>}{\sqrt{(<r_i(t)^2> - <r_i(t)>^2)(<r_j(t)^2> - <r_j(t)>^2)}}$, where i and j denote two stocks and $r_i(t) = \log p_i(t) - \log p_i(t-\tau)$. Here we use the closing price for $p_i(t)$, and also we set $\tau = 1$, in which case r is called the *log return* of the closing price. Thus if $p_i(t)$ is the closing price for the t^{th} day, then $p_i(t-1)$ is the closing price for the day before, i.e., the the $(t-1)^{th}$ day. The average $< \cdot >$ is a temporal average on the investigated time regime (for our data set its over 3 years). By definition, the value of ρ_{ij} ranges from -1 to 1. Can you explain why log return is a good measurement (do some Googling)?

2. **Constructing Correlation Graphs:** We can view every stock as a node in a graph, and the length of the link connecting two different stock return time series i, j is denoted by $d_{ij} = \sqrt{2(1-\rho_{ij})}$ for $i \neq j$. Plot the histogram of d_{ij}'s. Now construct a weighted graph G with adjacency matrix $D = [d_{ij}]$.

3. **Finding Minimum Spanning Trees (MSTs) for the Correlation Graphs:** Compute a minimum spanning tree (MST) for the correlation graph. Can you observe any particular pattern in the MST? Each stock can be categorized into a sector, which can be found from `Name_sector.csv` file. Plot the MST and color-code the nodes based on sectors. Do you see any pattern in this colorful graph? Interpret your observations.

4. **Evaluating Sector Clustering in MSTs:** Assume we want to predict the market sector of an unknown stock based on the MST

you just found. One might be interested in doing so just based on the immediate neighbors of a stock in the MST. Evaluate the performance of such a method as follows:

$$\alpha = \frac{1}{|V|} \sum_{v_i \in V} P(v_i \in S_i)$$

where S_i is the sector for the node i, and denoting neighbors of v_i by N_i, $P(v_i \in S_i) = \frac{|\{j | v_j \in N_i, S_j = S_i\}|}{|\{j | v_j \in N_i\}|}$.
Compare the above metric with the case where each MST node is assigned a random sector. That is, the probability a node is assigned the sector S_i is $\frac{(\text{\# of stocks with label } S_i)}{\text{Total \# of stocks}}$.

5. **Δ-Traveling Salesman Problem:** Determine if the triangle inequality holds for the fully connected graph G. Show your methodology. Now, we want to find an approximation algorithm for the traveling salesman problem (TSP) on G. Apply the algorithm described in the class (also see Chapter 17 in the recommended textbook authored by Papadimitriou and Steiglitz), which involves duplicating the edges in the MST you determined in the previous part. Then find an approximate tour for the TSP. Can you give a guarantee on the globally optimality of your solution?
For determining an Eulerian Cycle, you may want to use this code for your reference.

   Bonus Points: Use a general TSP solving package that does not exploit any patterns in the distances to solve the problem directly. Do you get a better tour than the approximation algorithm?

6. **Constructing Correlation Graphs for weekly data:** In the first part, we used daily closing prices for stocks to compute returns. Now, sample the stock data weekly on Mondays, and then calculate ρ_{ij} based on weekly data. Determine the related MST and compare the two results: based on daily and weekly stock prices. Here we ignore the holidays on Mondays, since there is no data available.

7. **Modifying Correlations:** Plot the histogram of ρ_{ij}'s from daily data. Then, set all ρ_{ij}'s larger than 0.3 as -1; for $\rho_{ij} \leq 0.3$, keep the

original value. Calculate $d_{ij} = \sqrt{2(1-\rho_{ij})}$. Construct the graph and run MST. The structures that you should find in the MST's for the unmodified correlations (if you did not make any mistakes) are called Vine Clusters. Do you see vine cluster patterns anymore with modified correlations? Can you explain why?

8. **A Generative Model for Vine Cluster Graphs [Bonus]:** The structures that you should find in the MST's (if you did not make any mistakes) are called Vine Clusters. One way to find vine clusters in the MST of graph G is through fitting a generative model. Can you come up with a generative model for vine cluster graphs?