Urban Computing-Saket Vishwasrao

1. (a) The best source of information that can be asked from WMATA is the ticketing inforamtion of all the passengers.This will help decide the volume of people that travel from station A to station B at any given time where A and B are any two stations on the metro map.

   To decide the optimum possible routes for the given network we can make a bipartite graph where all the junction stations should form one set(set A).The other set(let us call it set B) includes nodes such that each node represents a set of station between two junctions or between an edge point and a junction.For instance all the stations between Vienna and East Falls Church excluding this two stations are represented by one node.Every node in this set which is represents stations between and end point and junction has a degree one while every node that represents stations between two junctions has a degree 2(connected to 2 junction stations). Based on the the volume information that the tickets provide(we can determine the total inflow and outflow of people on one node of set A by summing up the people traffic of all stations on that node),we find a pair of nodes that have the highest traffic between them i.e maximum number of people want to travel between the stations represented by the two nodes.Then we find a path from the above graph that connects the two stations.While choosing the path if we have to choose between two sets of station we select the set(node) with higher volume of people.We basically want to offer services to majority of the people who want to travel between two stations such that they have to change a minimum number of trains.Also a constraint may be added based on the number of paths that can pass between two junction stations

   (b) For express trains, let us consider one particular route.We assign weights to each station on that route based on the total volume of people the station handles(inflow +outflow).So basically the station that most number of people use gets the most weight.Then based on the weight we select greatest k stations from a set of n stations along that route and make the trian stop at those stations.The value of k can be optimised based on the percent of population these k stations serve.For instance if 4 stations in a route of 10 stations handle more than 90 percent of the traffic it makes sense to stop trains at these stations.

2. (a) By limiting the spread of disease,we are trying to find the time threshold for the disease to stop spreading and minimize the number of nodes affected in that time period

   (b) One way to limit the spread of disease is to identify the nodes with highest connections i.e the ones with highest degrees and immunise them beacause the ones with highest connections are most likely to spread disease faster.We can use to degree distribution to find the nodes and to find the most central node based on the degree distribution and immunize those first.

   (c) Another way is to look at other measures such as betweeness centrality,closeness and identify nodes that are most central and try to remove those.

   (d) Mathematically we are trying to find a set of nodes such that when removed the spectral radius of the graph falls below a certain threshold.If that happens it is proved that the epidemic dies in logarithmic time.

3. For optimum sensor placement first find the nodes that have the highest degree and place sensor on those nodes.If the entire graph is not covered,place sensor on the nodes having second highest degree distribution and so on.Repeat this process till the entire graph is covered.

4. (a) Figure 1,2,3 show the road degree distribution of the road networks

   (b) No the networks are not connected.Yes the networks can be tested for connectedness.One way is depth first search.In depth first search we randomly select a node from the graph as the root node.Then recursively visit each of its neighbours and mark the visited nodes.

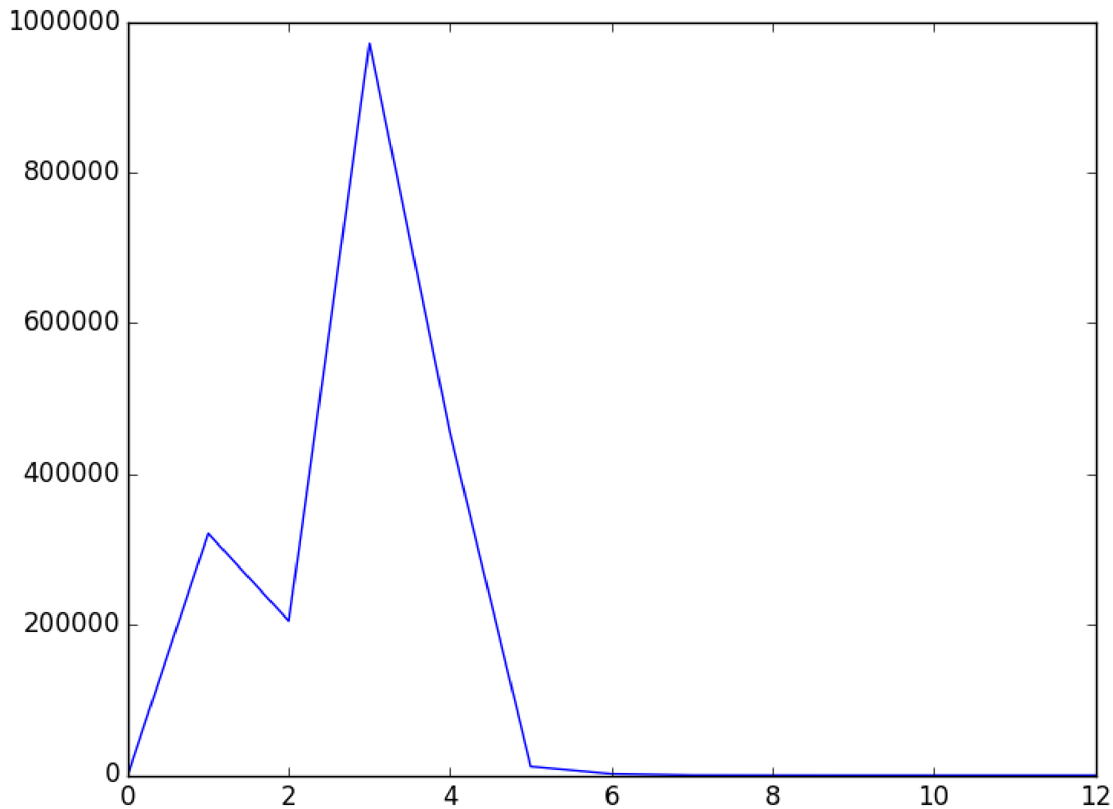

Figure 1: California Road Network

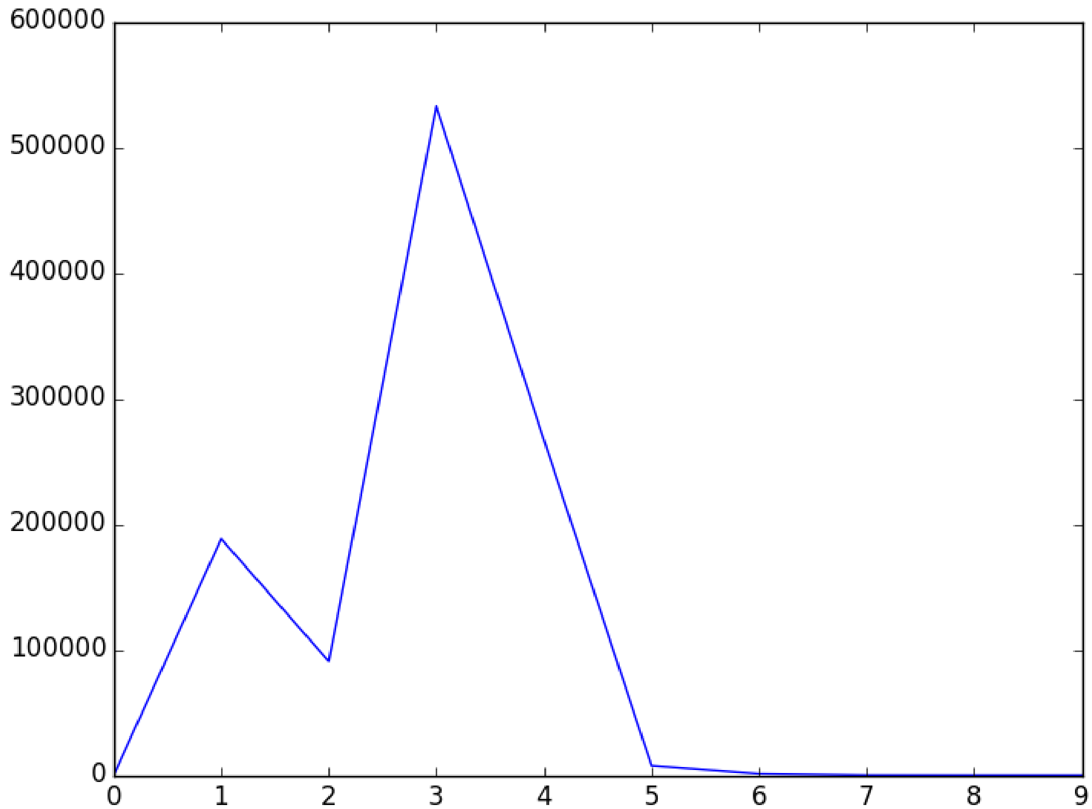

Figure 2: PA Road Network

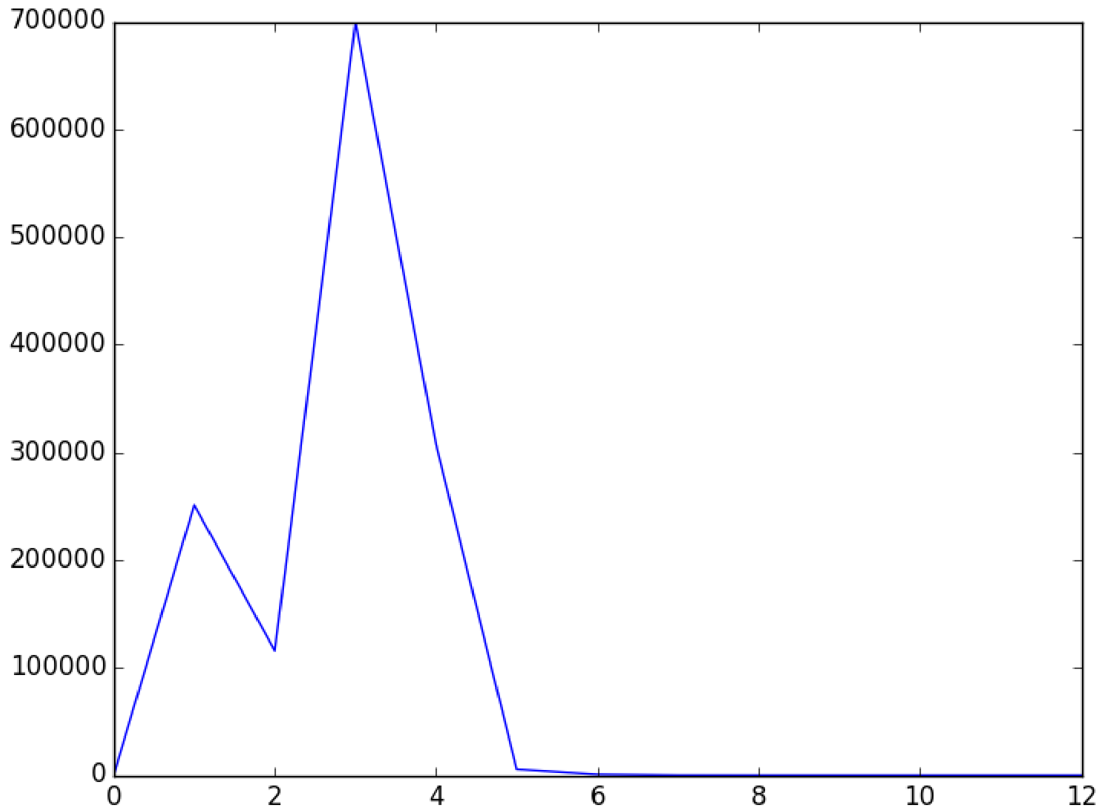

Figure 3: Texas Road Network

procedure DFS(G,v):
mark node v
for all neigbours in G.adjacentnodes(v) do
if vertex w is not marked then
recursively call DFS(G,w)

(c) The least number of nodes to be removed so that the network is not connected is to remove the nodes with highest degree centrality one by one and test for connectedness.

(d) California
Open Wedges:5874414,Closed Wedges: 120676,Total Wedges:5995090

Philadelphia
Open Wedges:3323517,Closed Wedges: 67150,Total Wedges:3390667

Texas
Open Wedges:4045156,Closed Wedges: 82869,Total Wedges:4128025

(e) wedge ratio
California:0.8366
Philadelphia:0.8269
Texas:0.8180
The wedge ratio indicates fraction of nodes through which multiple nodes pass.

5. Algorithm written.Code execution taking too long

6. (a) The number of nodes that are reachable in exactly t steps are $k(k-1)^{t-1}$

(b) The number of nodes except the leaf nodes are

$$\sum_{i=1}^{t} k(k-1)^{t-1} + 1 = \frac{k}{k-2}((k-1)^t - 1) + 1 \tag{1}$$

where $(t+1)$ is the number of steps needed to reach the leaf node from the central node.
Solving for t

$$n = \frac{k}{k-2}((k-1)^t - 1) + 1 + k(k-1)^t \tag{2}$$

$$\implies (n-1) = \left(\frac{k}{k-2} + k\right)(k-1)^t - \frac{k}{k-2} \tag{3}$$

$$\implies (n-1) = \frac{k(k-1)}{k-2}(k-1)^t - \frac{k}{k-2} \tag{4}$$

$$\implies \frac{(n-1)(k-2)}{k} = (k-1)^{t+1} - 1 \tag{5}$$

$$\implies t = \frac{log\left(\frac{(n-1)(k-2)}{k} + 1\right)}{log(k-1)} - 1 \tag{6}$$

Hence the degree is 1 for $k(k-1)^t$ leaf nodes and k for rest of the $n - k(k-1)^t$ nodes where t is represented by equation 6.

(c) The diameter in terms of is $2t$ where t given by equation 6

(d) Cayley trees are useful in urban computing because they can be used to represent the distribution network of an utility in a city for example electricity.The central node represents the main power generation station while the other nodes can be viewed as a substation.A sample electricity distribution network in a city can be represented by multiple cayley trees where each tree represents the power station and its distribution network.Same can be said about water distribution etc.