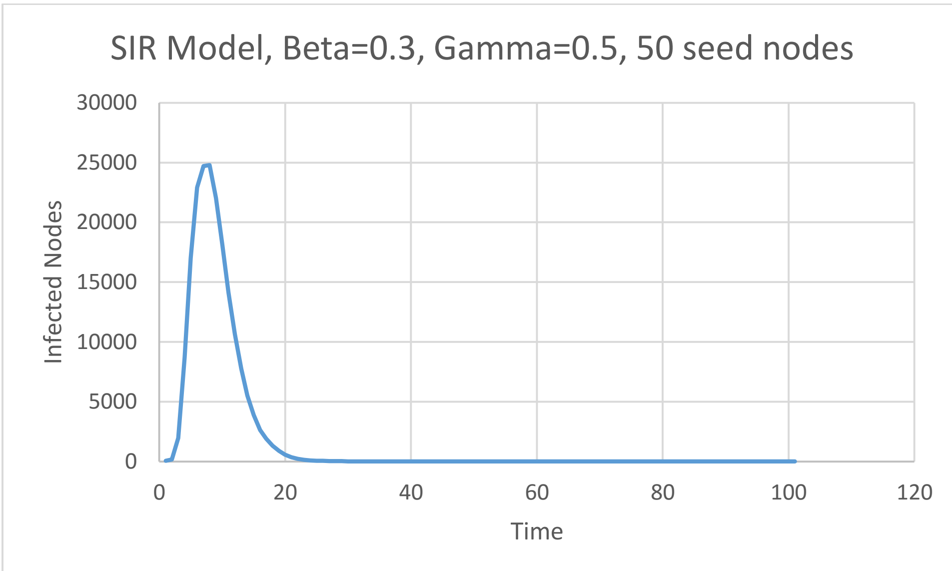
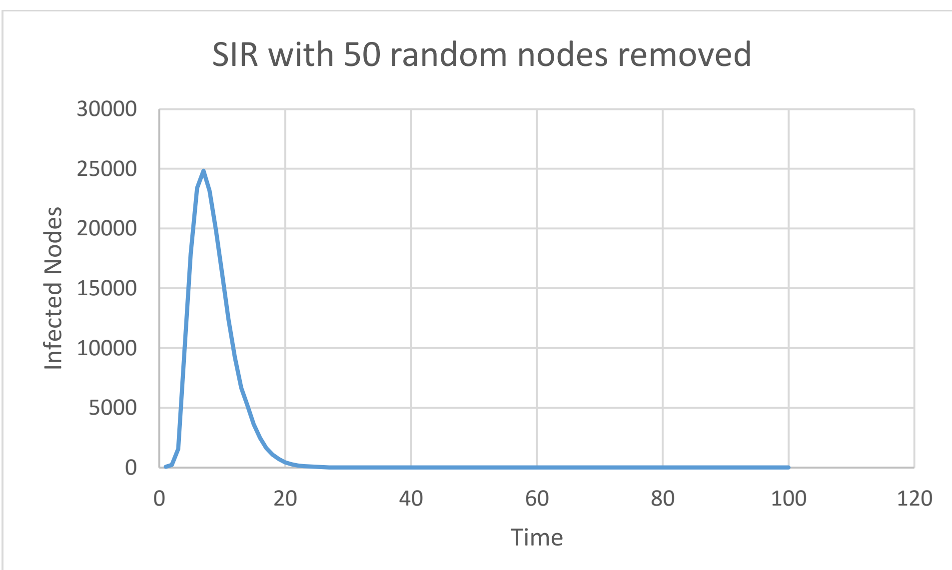
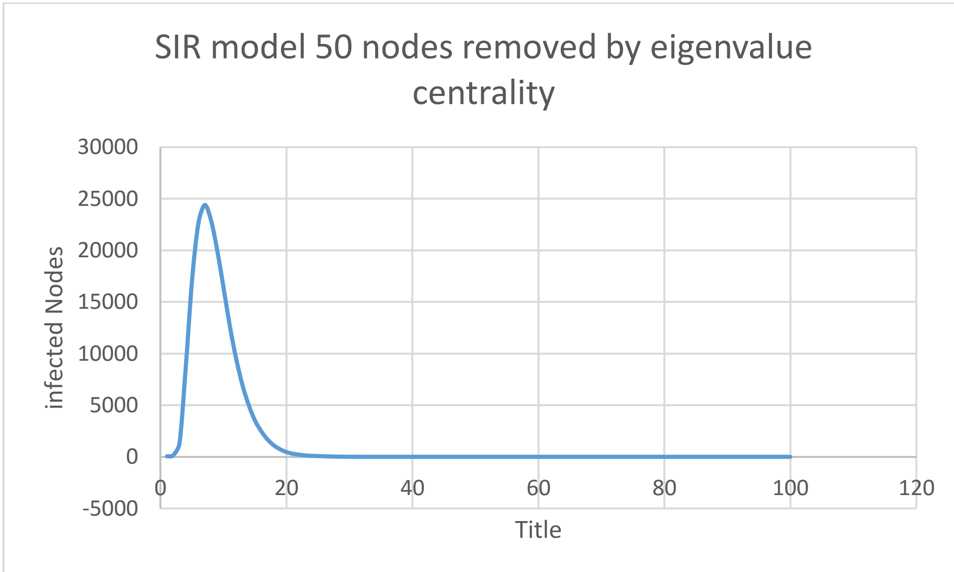
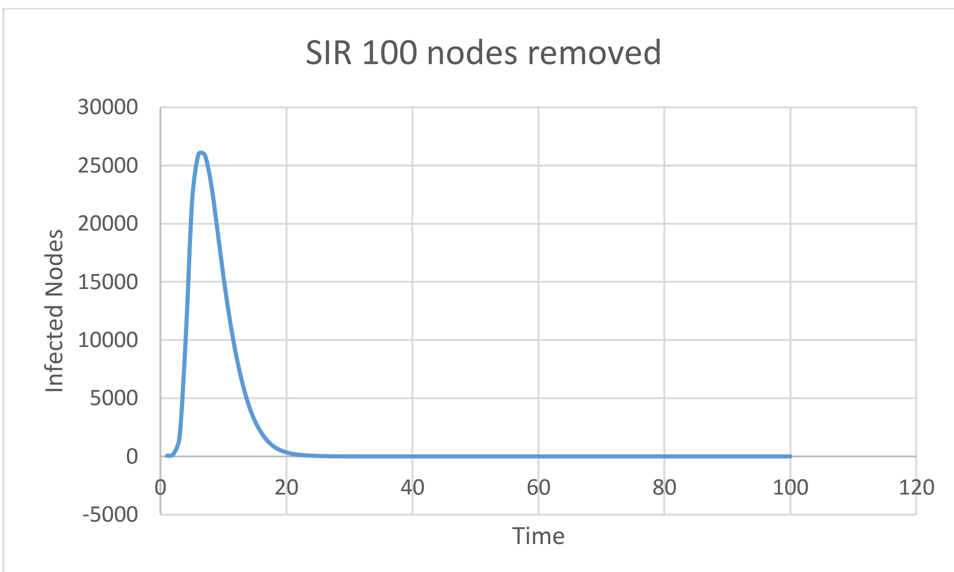
Urban Computing Assignment 3-Saket Vishwasrao

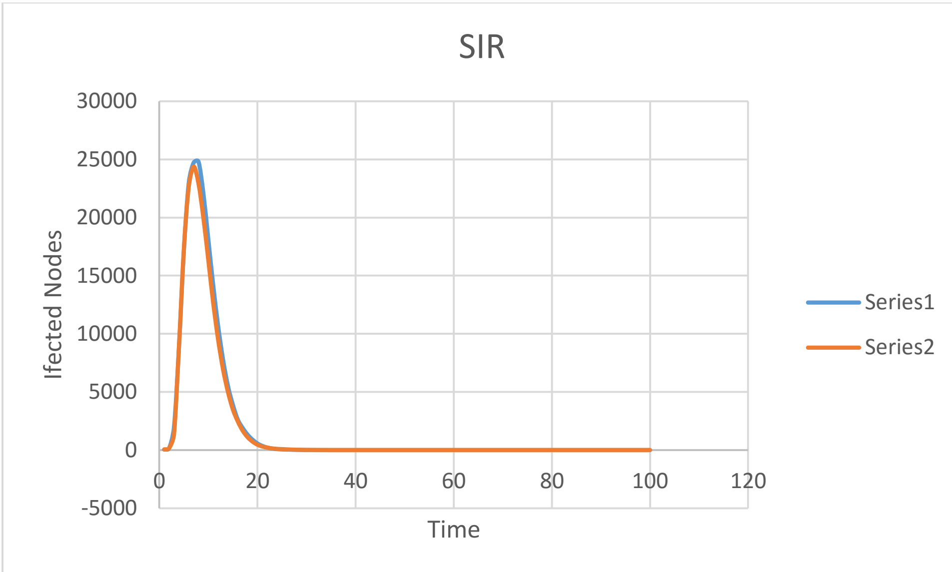
1. The normal SIR model takes longer time for stabilizing to zero. We can use the friendship paradox for picking the nodes to be removed. As per the friendship paradox, most people have fewer friends than their friends have. This implies that friends of randomly selected individuals have higher average centrality. Thus removing randomly selected nodes, we see that the the infection spreads slower than the earlier case as the number of susceptible people decrease due to the removal of highly connected individuals. Thus, the epidemic spread is reduced.





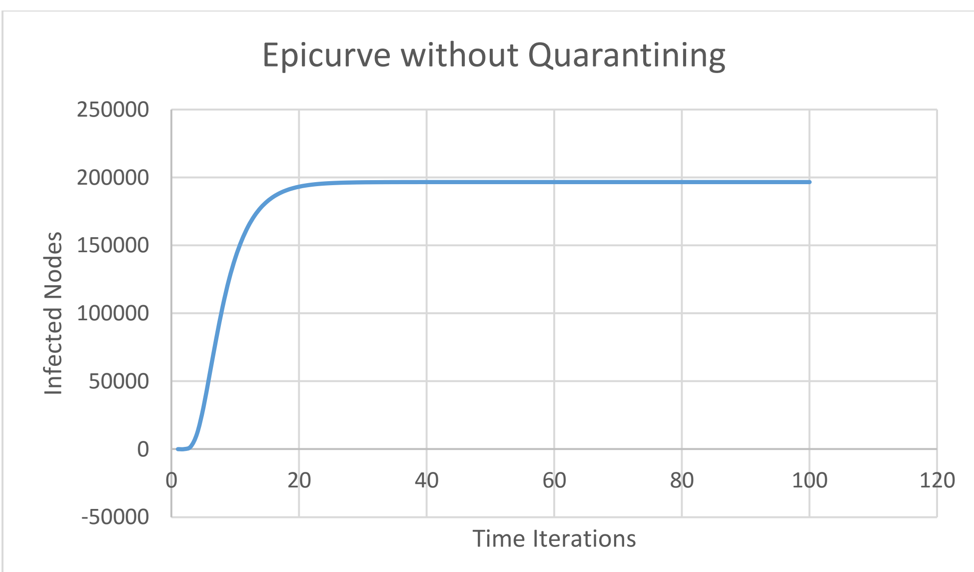

SIR model 50 nodes removed by eigenvalue centrality
infected Nodes
30000
25000
20000
15000
10000
5000
0
-5000
0
20
40
60
80
100
120
Title

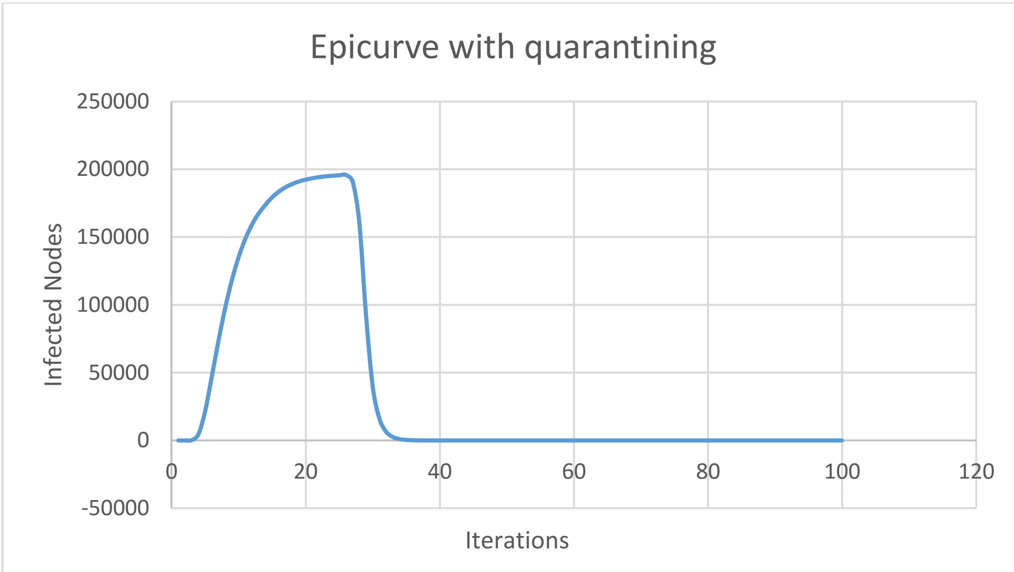

SIR 100 nodes removed
Infected Nodes
30000
25000
20000
15000
10000
5000
0
-5000
0
20
40
60
80
100
120
Time

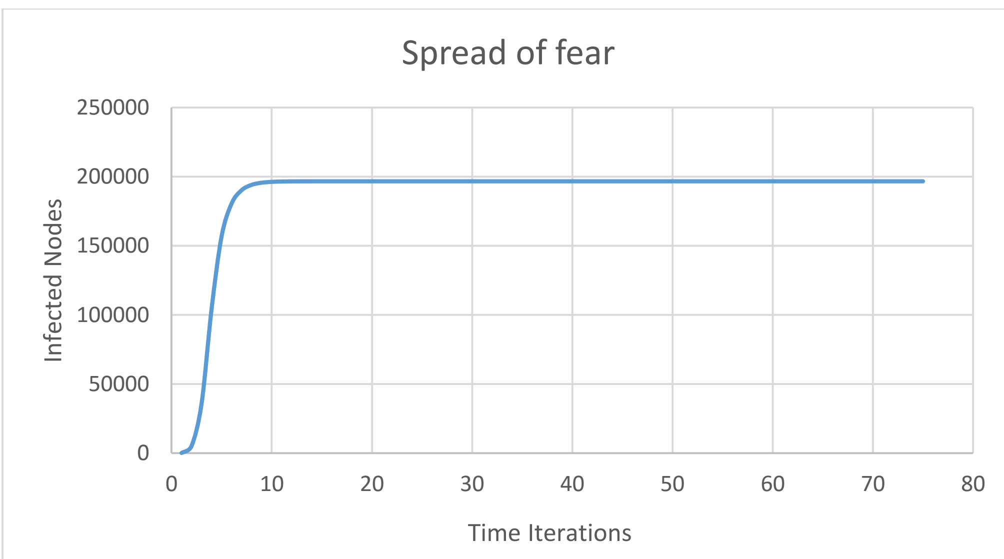


Removing nodes does not affect the SIR curve. It is almost the same when 50 nodes are removed. This may be due to the fact that the graph nodes are highly connected and it does not matter if we remove 50 nodes, the curve remains the same.

B)







The Epicurve of I won't affect the other fear network because the fear spread is not dependent on disease spread however the vice versa i.e the disease spread is dependent on fear spread. Yes the quarantining approach limits the spread of disease. However, it can be observed from the graph that the till time=25, more than 95% of the nodes are infected. It would have been better if we started quarantining earlier.

I use eigenvector centrality for the seed nodes to maximize fear spread. The more the nodes to whom fear is spread is quarantined, the more faster the number of infected nodes falls. Thus quarantining helps in preventing epidemic spread.

2. Solutions
    a. We first do a region based clustering on the trajectory patterns. For instance we can partition repetitively between X and Y axes such that the MDL cost decreases. We basically get a set of individuals who have been within a same region once we classify their trajectories. Then we use algorithm similar to periodic pattern mining on the sub trajectories of individuals in a particular region classified earlier. We do this by taking the trajectory points of all individual trajectories in a region within a time interval t and performing clustering on them. We do this for all time intervals. The number of individuals whose trajectories fall in one cluster of a particular time interval t have high probability of spreading a disease. We can thus create a contact graph based on this information. Once we have a contact graph we can track global spread of disease using one of the disease spread model.
    b. We again first perform a region based clustering to find the number of trajectories passing through a particular region. Within a particular region we try to find subsequences that occur the highest using say using Prefix Span. This gives us the most used routes in a particular region. We then performing clustering of the subsequences based on temporal data to find the net volume of traffic on a particular route. Once we have this information, we can easily design policy to optimize the traffic flow.
    c. We can use probabilistic clustering or density based clustering to group similar trajectories of hurricanes, tornadoes etc. Thus given the past data, we can easily predict which cluster a new tornado/ hurricane will fall.

3. (a) Solution

    i. Supervised learning requires training on some labelled data. In energy measurement devices it is expensive in both labor and cost to create the training dataset. Such a problem is not present in unsupervised dataset.
    ii. In supervised learning we assume a fixed number of classes (devices) and create a model for their corresponding signatures. However this is mostly not the case in real scenarios because there are often unknown loads. Unsupervised learning can adapt in such situations.

   (b)

    i. We first find a steady state plot of power for our analysis
    ii. Generally a refrigerator is used all day in a household, whereas a dryer is used only during a particular time in week. We first have to do a temporal mining to find similar

sequences (episodes) in the curve. The episodes corresponding to refrigerator will be more frequent than the dryer as refrigerator is used everyday.

iii. We can then use a supervised learning algorithm (naïve bayes) if we know the model for refrigerators and dryers or we can use unsupervised learning like HMM to predict the device corresponding to a particular episode. Such a learning algorithm may not be necessary if the patterns are very distinguishable.