Department Of Computer Science

PAGERANK, DIFFUSION & CASACDING OF COVID-19 ON THE NETWORK OF ALL STATES/UT IN INDIA

SUBMITTED BY:

SAWON BHATTACHARYA

(19370045)

MSc 1st YEAR

Abstract:

Basically Epidemic models are used to explain the spread of the contagious diseases. Here we consider a model of diffusion where we will discuss about the certain parameters and how they play a key role in spreading the viruses like COVID-19. Also we will show how fast a disease like COVID-19 can cascade through the network and create a huge damage on human life based on the connectivity and the concept of the key nodes. Before discussing about diffusion firstly we will discuss about the Page Rank algorithm. We will implement these concepts on a real life dataset. After implementing the concepts in python we will visualise the results and draw a conclusion.

Key Words

Social network, PageRank, epidemic, diffusion, cascading

1. Introduction

In this current situation COVID-19 has create a havoc impact all over the world. Here we have considered only the cases in India. Recently an event called “Tablighi Jamaat” took place in Delhi (the capital of India) has become a huge factor in increasing the numbers of people infected by this virus. Here we have created a network of all states and union territories based upon the number of cases found in the dataset.

Every node in the network is a state like Delhi, West Bengal Tamil Nadu etc. Edge between a pair of state represents a direct connection between them. By analysing these network we will calculate the pagerank of each state.

After pagerank we will discuss how the event “Tablighi Jamaat” becomes a huge factor in diffusing the COVID-19 virus all over the India. Also analyse the infectiousness of the corona virus and how fast it has spread through the network by implementing Cascading in the network.

At last we will compare the number of confirmed cases (State wise) based upon the date of “Tablighi Jamaat”.

2. Dataset

My datasets come from an acclaimed website “Kaggle”. **Kaggle** is an online community of data scientists and machine learning practitioners. Kaggle allows users to find and publish data sets, explore and build models in a web-based data-science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.

Context

a. January 30: the first case of the COVID-19 in India was reported, originating from China.
b. March 14: Central government declares COVID-19 a 'notified disaster'
c. March 15: The number of positive cases crosses 100
d. March 23: People participated in “Tablighi Jamaat” started flying from Delhi to several states and cities of India. (For further details please check Reference point, there I have attached the link of the datasets.)

I have used this dataset to create my own dataset. One csv file contains the links or paths between states or union territories. If there is a path then it will be represented by 1 else it will be 0.From this file I have created the network to analyse the

PageRank, Diffusion and Cascading. Other one contains the number of total confirmed cases of each states or union territories.

3. Pagerank:

Page Rank is extensively used for ranking web pages in order of relevance by mostly all search engines world-wide. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

The basic definition of PageRank: Intuitively, we can think of PageRank as a kind of "fluid" that circulates through the network, passing from node to node across edges, and pooling at the nodes that are the most important. Specifically, PageRank is computed as follows:

1. In a network with n nodes, we assign all nodes the same initial PageRank, set to be 1=n.
2. We choose a number of steps k.
3. We then perform a sequence of k updates to the PageRank values, using the following rule for each update:

Basic PageRank Update Rule: Each page divides its current PageRank equally across its out-going links, and passes these equal shares to the pages it points to. (If a page has no out-going links, it passes all its current PageRank to itself.) Each page updates its new PageRank to be the sum of the shares it receives.

Mathematical concept behind the page rank algorithm:

In 1998, the founder of Google search engine, Larry Page and Sergey Brin invented the Page Rank Algorithm to quantize the importance of millions of web pages comprising the World Wide Web (WWW).

The basic concept of PageRank is that the importance of a page is directly proportional to the number of web pages linking to that page. So Page Rank algorithm considers a page more important if large number of other web pages are linking to that page or if links are coming from some of most important and popular web pages. Page Rank of whole website is not valid because page rank is associated with every web page on the web. Page Rank of a web page X is calculated by the page rank of those pages that links to page X using formula given below:

$PR\ X = 1-d + [PR(Y1)/C(Y1) + \ldots\ldots + PR(Yn)/C(Yn)]$

Where,

PR (X) = Page Rank of web page X,

PR (Yi) = Page Rank of pages Yi that links to a web page X

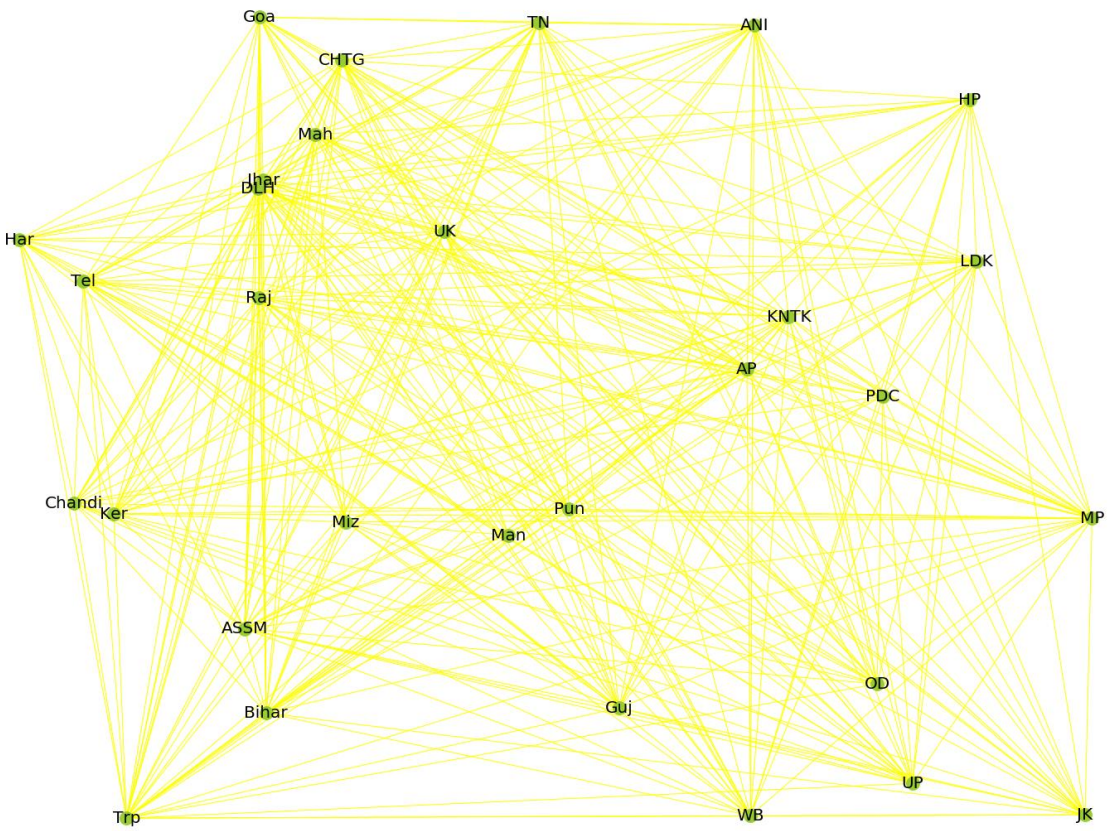
C (Yi) = Number of outbound links on web page Yi

d = Damping Factor (value between 0 and 1, but usually value is 0.85)

Repeat the above step involving the calculation of page rank until two consecutive same values are obtained.

Though PageRank algorithm is used to find the rank or importance of the web pages, we have used this concept in my network to derive the most important node based on the connectivity of the node through the network. By this we can say that if a virus spread through the network then the node with highest pagerank can play a crucial role.

From this network (img1) mentioned below, we have calculated the pagerank of each node where each node represents a state or union territory.



Img1: Network of states and union territories in India

No	State/Union Territory	PageRank
1	ANI	0.0325420559290275
2	AP	0.040474610866908914
3	ASSM	0.039226597929005735
4	Bihar	0.033942860865118475
5	Chandi	0.03006269719650649
6	CHTG	0.03922100265230109
7	DLH	0.040584905984835995
8	Goa	0.029909662627653863
9	Guj	0.03259358781940482
10	Har	0.0288926349684197
11	HP	0.027649672542665846
12	JK	0.030100901761055295
13	Jhar	0.03894961406897551
14	KNTK	0.03640141972896503
15	Ker	0.032621543297413624
16	LDK	0.03135250945126094
17	MP	0.03894961406897551
18	Mah	0.03551708249242784
19	Man	0.0149773762496505 23
20	Miz	0.02777167951808753
21	OD	0.0288236701826375
22	PDC	0.025209519619062995
23	Pun	0.032589827044835255
24	Raj	0.03639985758984221
25	TN	0.03634191578550636
26	Tel	0.03668065723812436
27	Trp	0.04053662161062194
28	UP	0.03265126964032743
29	UK	0.03639150874309246
30	WB	0.032633122527289025

DLH has the highest Pagerank: 0.040584905984835995
Hence Delhi can be most dangerous state in terms of becoming a vector.

Img2: pagerank of states and union territories in India

4. Diffusion & Cascading

How something spreads across the network is a key question we will be interested in asking when analysing many different types of networks the classical example of this being the diffusion of some disease through a population. More formally we call this spreading on a network propagation or diffusion. How diffusion happens and how long it takes is defined by number of parameters we will just list the primary factors are involved here before looking at them individually.

Primary Parameters for Diffusion:

a. **Infectiousness metric**. By infectiousness we mean, something that is likely to spread or influence others in a rapid manner irrespective of the type of network that is spreading on. We can quantify this in terms of how contagious the diseases is

b. **Resistance.** By resistance we mean, how resistant the nodes in the network are to spreading of this phenomena. Also we can add the time model with this point. It captures how many nodes may be infected for only a brief period of time before recovery.

c. **Topology.** Topology helps us to understand that how something is likely to spread across the network. The primary factor here is simply the overall degree of connectivity to the network. Obviously the more connected it is the fastest it is the faster something should spread across it.
We also need to analyse the degree distribution to understand how centralised the network is. As centralized networks with major hubs enable rapid local and global diffusion

d. **Dissemination Strategy.** We also need to look for whether the dissemination is random or strategic. This means that whether there is some logic behind the promotion and dissemination aimed as strategically effecting nodes that have a high degree of connectivity and thus enabling a more rapid diffusion.
.

Modelling Diffusion through a Network
We build our model for the diffusion of a new behaviour in terms of a more basic, underlying model of individual decision-making: as individuals make decisions based on the choices of their neighbours, a particular pattern of behaviour can begin to spread across the links of the network.

To illustrate our point, consider the basic game-theoretic diffusion model proposed in [7]. Consider a graph G in which the nodes are the individuals in the population and there is an edge (i, j) if i and j can interact with each other. Each node has a choice between two possible behaviours labelled A and B. The payoffs are defined as follows:

1. if i and j both adopt behaviour A, they each get a payoff of $a > 0$;
2. if they both adopt B, they each get a payoff of $b > 0$; and
3. if they adopt opposite behaviours, they each get a payoff of 0.

In this situation j's selection of behaviour will depend upon the following: suppose that some of its neighbours adopt A, and some adopt B; what should j do in order to maximize its payoff? This clearly depends on the relative number of neighbours doing each, and on the relation between the payoff values a and b. With a little bit of algebra, we can make up a decision rule for j quite easily, as follows. Suppose that a p fraction of j's neighbours have behaviour A, and a (1- p) fraction have behaviour B;

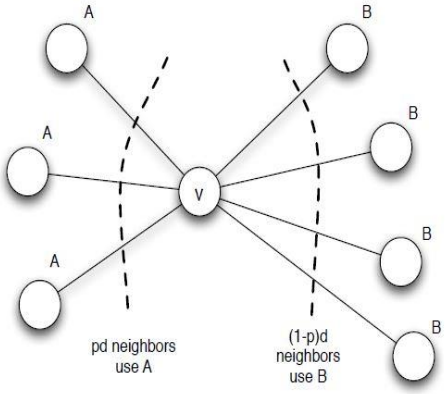


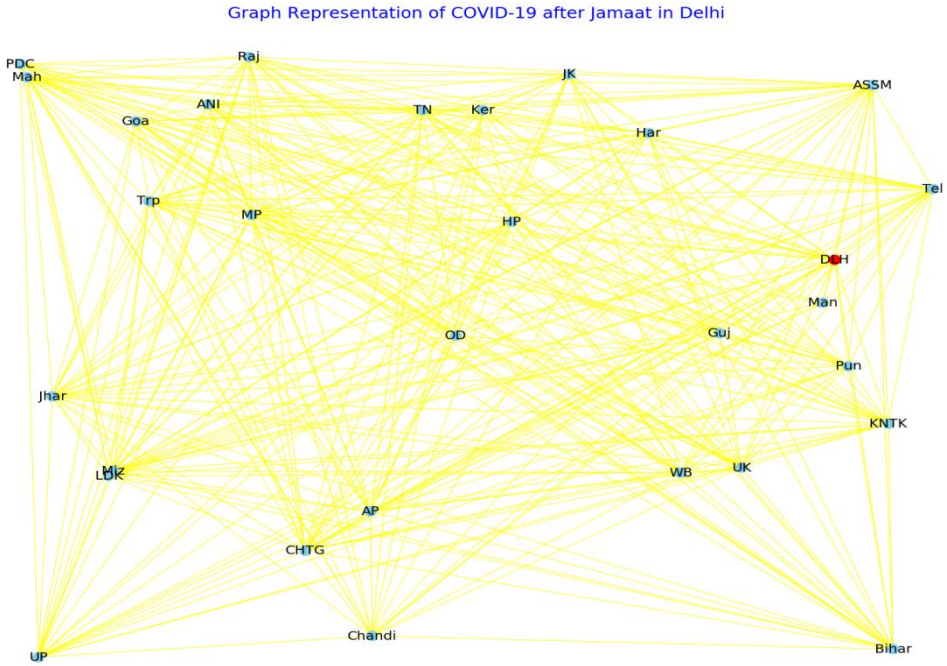
Figure 19.2: v must choose between behavior A and behavior B, based on what its neighbors are doing.

Img3: j must choose between A and behaviour B, based on what its neighbours are doing.

that is, if j has d neighbours, then pd adopt A and (1 - p)*d adopt B, as shown in Figure 19.2. So if v chooses A, it gets a payoff of pda, and if it chooses B, it gets a payoff of (1 - p)*db. Thus, A is the better choice if pda > (1- p)*db; or, rearranging terms, if

p >= b/a + b

We'll use q to denote this expression on the right-hand side. This inequality describes a very simple threshold rule: it says that if at least a q = b/(a+b) fraction of your neighbours follow behaviour A, then you should too. And it makes sense intuitively: when q is small, then A is the much more enticing behaviour, and it only takes a small fraction of your neighbours engaging in A for you to do so as well. On the other hand, if q is large, then the opposite holds: B is the attractive behaviour, and you need a lot of your friends to engage in A before you switch to A.

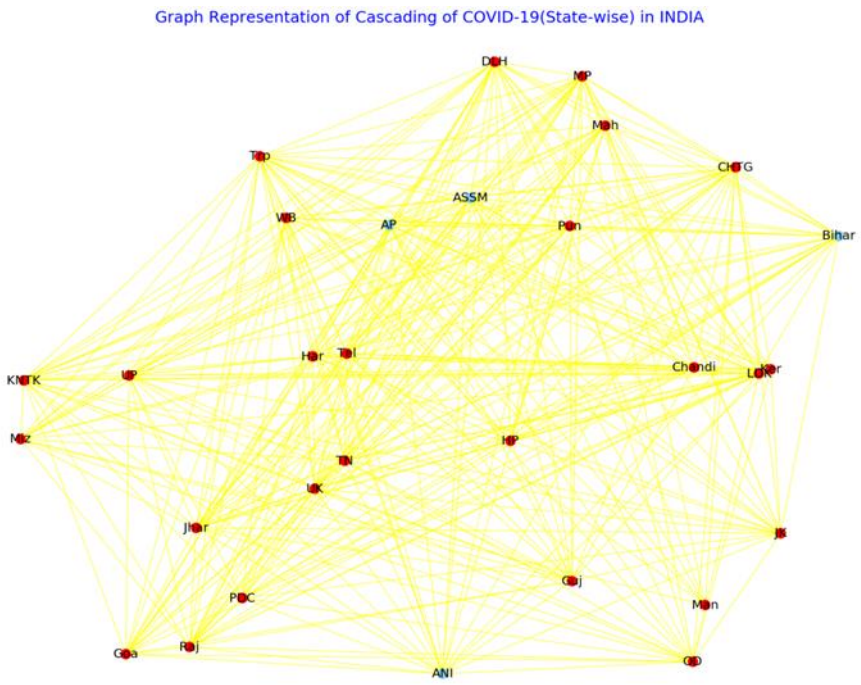


Img 4: Cascading started from Delhi (red node)

Cascading Behaviour: In any network, there are two obvious equilibria to this network-wide coordination game: one in which everyone adopts A, and another in which everyone adopts B. Where A is used by the initial adopter and B is already there in the network. Now there may be another scenario where A and B coexists in the network. From these scenarios we would like to distinguish two different possibilities:

(I) that the cascade runs for a while but stops while there are still nodes using B, or

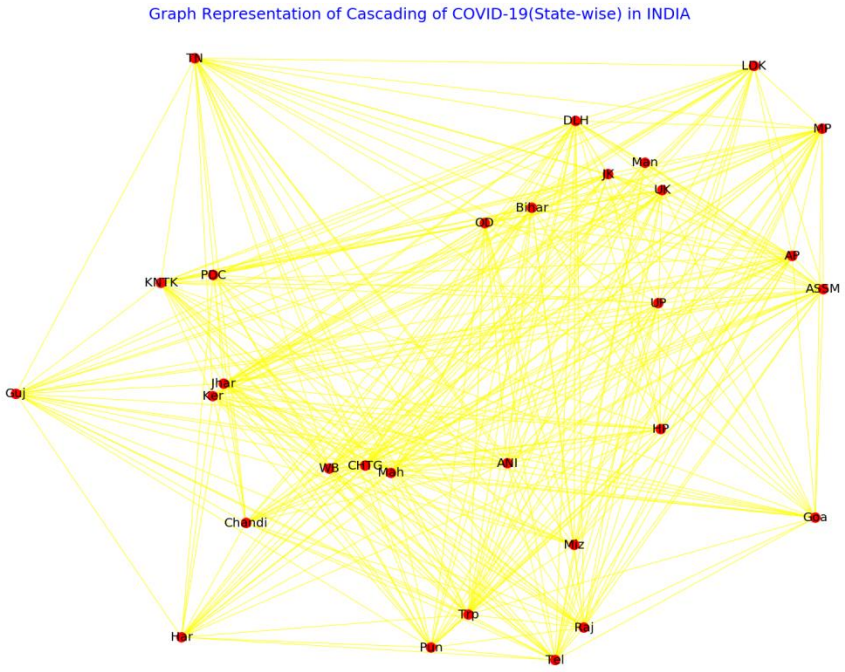
(II) that there is a complete cascade, in which every node in the network switches to A.



Img 5: Cascading step 2 (red nodes have increased)

We consider the following type of situation. Suppose that everyone in the network is initially using B as default behaviour. Then, a small set of \initial adopters" all decide to use A. We will assume that the initial adopters have switched to A for some reason outside the definition of the coordination game - they have somehow switched due to a belief in A's superiority, rather than by following payoffs - but we'll assume that all other nodes continue to evaluate their payoffs using the coordination game. Given the fact that the initial adopters are now using A, some of their neighbours may decide to switch to A as well, and then some of their neighbours might, and so forth, in a potentially cascading fashion. When does this result in every node in the entire network eventually switching over to A? And when this isn't the result, what causes the spread of A to stop? Clearly the answer will depend on the network structure, the choice of initial adopters, and the value of the threshold q that nodes use for deciding whether to switch to A.

The above discussion describes the full model. An initial set of nodes adopts A while everyone else adopts B. Time then runs forward in unit steps; in each step, each node uses the threshold rule to decide whether to switch from B to A.1 The process stops either when every node has switched to A, or when we reach a step where no node wants to switch, at which point things have stabilized on coexistence between A and B.



Img 6: Cascading step 3 (red nodes have increased)

In this network all nodes initially are not that much affected and symbolized by “sky-blue” colour, except the node “DLH” (Delhi). That node with “red” colour represents the event “Jamaat”. After that event the novel virus had increased rapidly from Delhi across the network. We have kept the payoff of each “red” node = 100 based upon the parameter infectiousness of the virus. The payoff of each “sky-blue” node =5.

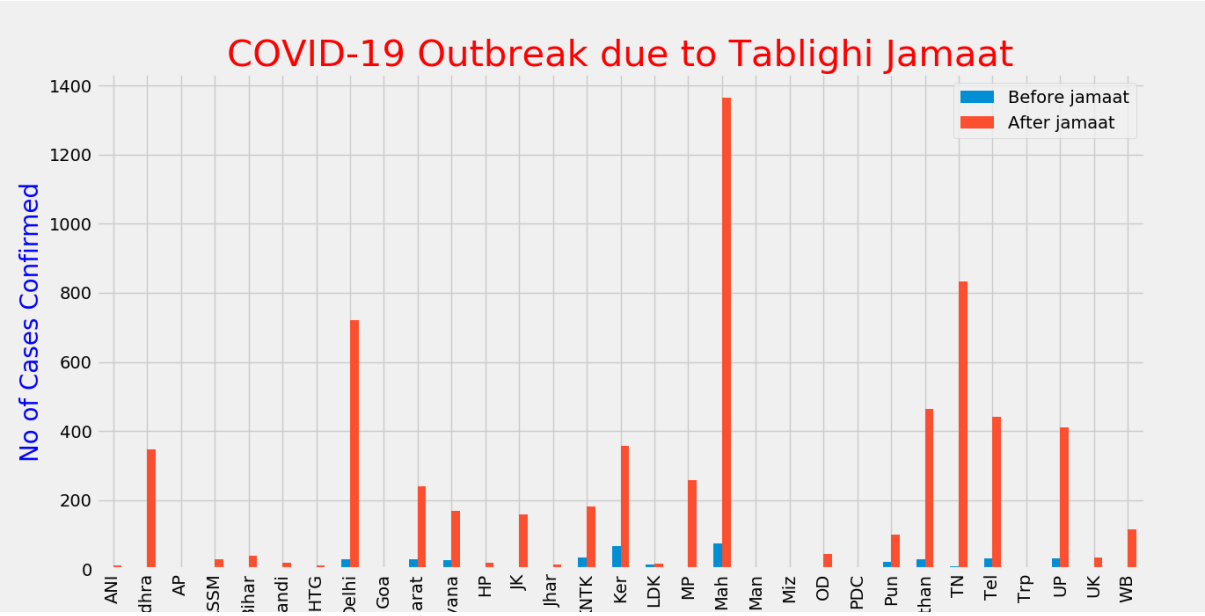
From the images we can see that Cascading behaviour had started from Img4 and in the Img6 we got the full Cascade model. That depicts that just in between 2 iterations the virus had diffused through the network. So from here we can say that not only because the payoff of red node was high but also the structure and connectivity of the network were also the other factors.

5. Comparison of data

Here basically I have compared two dates of the total number of confirmed cases. One is the before “Jamaat” and another one is after “Jamaat”.From the graph we can see the epic difference in the total number of confirmed cases. Here I am giving some number from my dataset file “ba_values.csv”.

Name of State / UT	Before Jamaat	After Jamaat
Mah	74	1364
TN	9	834
Delhi	29	720
Rajasthan	28	463
Tel	32	442
UP	31	410
Ker	67	357
Andhra	7	348

Let's visualise the whole data set and see how badly the each state or Union Territory in India has been affected.



6. References

a. Networks, Crowds, and Markets: Reasoning about a Highly Connected World by, David Easley Dept. of Economics Cornell University & Jon Kleinberg Cornell University.

b. PageRank and random walks on graphs by, Fan Chung and Wenbo Zhao University of California, San Diego La Jolla, CA 92093, US

c. A Review Paper on Page Ranking Algorithms by, Sanjay* and Dharmender Kumar Department of Computer Science and Engineering, Guru Jambheshwar University of Science and Technology.

d. Diffusion and Cascading Behaviour in Random Networks by, Marc Lelarge INRIA-ENS Paris, France marc.lelarge@ens.fr.

e. Dataset: https://www.kaggle.com/imdevskp/covid19-corona-virus-india-dataset