Importance of contact tracing while predicting epidemics
and
Epidemic Localization in networks with high order structure

Report submitted in fulfillment of the requirements
for the Exploratory Project of

Second Year IDD

By

Neha Kumari

Under the guidance of
Dr. Hari Prabhat Gupta

Department of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY (BHU) VARANASI Varanasi 221005, India
May 2020

Dedicated to:

My parents, professor and everyone who helped and motivated me in successful completion of this report.

Declaration

I certify that

1. The work contained in this report is original and has been done by myself and the general supervision of my supervisor.

2. The work has not been submitted for any project.

3. Whenever I have used materials (data, theoretical analysis, results) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references.

4. Whenever I have quoted written materials from other sources, I have put them under quotation marks and given due credit to the sources by citing them and giving required details in the references.

Place: IIT (BHU) Varanasi
Date:- 05-06-2020

Neha Kumari.
(18074012)
(IDD Student)
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Certificate

This is to certify that the work contained in this report entitled "**Importance of contact tracing while predicting epidemics And Epidemic Localization in networks with high order structure**" being submitted by **Neha Kumari(18074012)** carried out in the Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, is a bona fide work of our supervision

Dr. Hari Prabhat Gupta
Department of Computer Science and Engineering,
Indian Institute of Technology (BHU) Varanasi,
Varanasi, INDIA 221005.

Place: IIT (BHU) Varanasi
Date: 05-06-2020

Acknowledgments

I would like to express my sincere gratitude to our respected Dr.Hari Prabhat Gupta sir,who was my mentor on an exploratory project. He helped me a lot at every stage during my project to clear my doubts.

Abstract

The subsequent outbreak of an epidemic has always led to sudden loss in human capital , world wide economy and deaths of large communities of people in a very short duration of time. If not taken care of properly, the epidemic can convert into pandemic within the blink of an eye.

This project outlines various factors on which the growth rate of pandemic depends.It also depicts what determines whether the epidemic will die out or eventually spread into the whole community. Various models have been proposed to study the growth rate of these epidemics and to find out how contact tracing can help us decrease the spread of the disease.

We will introduce a notation that will play a significant role in every type of epidemic model. That notation is R_0 . We will come to know from where does this notation come from and how does it affect the outbreak size of a pandemic. We will use network epidemiological study to find out what other factors apart from R_0 will help us reduce the pandemic outbreak.

we will demonstrate that there is a huge uncertainty in outbreak size without data on the heterogeneity in secondary infections for emerging pathogens like novel coronavirus.
Altogether ,this project highlights the critical need for contact tracing during emerging Infectious disease outbreaks and the need to look beyond R_0 when predicting epidemic size.

Lastly,We will also show there are sometimes mesoscopic localisations in epidemics.It is a phenomenon in which the epidemic is only concentrated around certain substructures in the network and not the entire network. We will discuss how the practical implications of blanket cancellation of events, school closures and social distancing has led to mesoscopic localisation.

In addition , we will show that mesoscopic localisations are there in epidemics will higher order interactions unlike the classic diffusion of standard epidemic models.
Unlike standard models of delocalized dynamics, epidemics in a localized phase can suddenly collapse and die out stochastically [2].

Contents:

List of symbols:

Symbol	Description
R_0	Basic Reproductive number.
$S(t)$	Fraction of susceptible population.
$I(t)$	Fraction of infected population.
$R(t)$	Fraction of recovered population.
β	Transmission rate.
γ	recovery rate.
G_0	Probability generating function.
$R\ (\infty)$	Outbreak size.
σ^2	Variance.
p_k	Probability of a node having k number of contacts.

Chapter 1

Introduction

1.1 Overview

For any epidemiological study, we first require the basic reproductive number of that epidemic.It is the most misapplied number in public health. It is the governing term for any epidemic. The basic reproductive number of any epidemic is the expected number of secondary infections that an infected person would cause. If R_0 =2, then one person is most likely to infect 2 other individuals ,those two can future infect 2 thereby increasing exponentially.

Mathematically,

$$R_0 = \frac{\beta}{\gamma}$$

Various epidemiological models have been devised to study patterns of growth of epidemiology and come to a conclusion on how to limit the outbreak size.
The models that have been devised so far are:

- SI model (susceptible infected)
- SIS model (susceptible Infected susceptible)
- SIR model (susceptible infected Recovered)

Mathematical expressions for R_0 in all the above models comes out to be the same.
The SIR model covers all aspects of the epidemics in a known susceptible population.

Infection equation:

$$\frac{dS(t)}{dt} = -\beta S(t)I(t)$$

$$\frac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t)$$

$$\frac{dR(t)}{dt} = \gamma I(t)$$

$$S(t) + I(t) + R(t) = 1$$

Solving the above differential equation , we can find out the fraction of susceptible, infected and recovered populations as well. There is no analytical solution for the above differential equation. However, this requires a deep knowledge of differential calculus because the rate of change of fraction of susceptible as well as infected population are varying in accordance with two independent variables.
Solving thoroughly these differential equations will yield the following results.

$$\frac{ds(t)}{dt} = -\beta s \frac{dr}{dt}\frac{1}{\gamma}$$

$$S = S_0 e^{-\frac{\beta}{\gamma}r}$$

$$\frac{dr}{dt} = \gamma(1 - r - s_0 e^{-\frac{\beta}{\gamma}r})$$

Limits: $t \to \infty, \ \frac{dr}{dt} = 0 \ , R(\infty) = constant$ ....(1)

At t=0, the fraction of recovered population is zero as everyone is susceptible and no one has recovered yet. So, $R(0) = 0, \ I(0) = i_0, \ S(0) = 1 - i_0 \simeq 1$

Therefore according to eq (1)

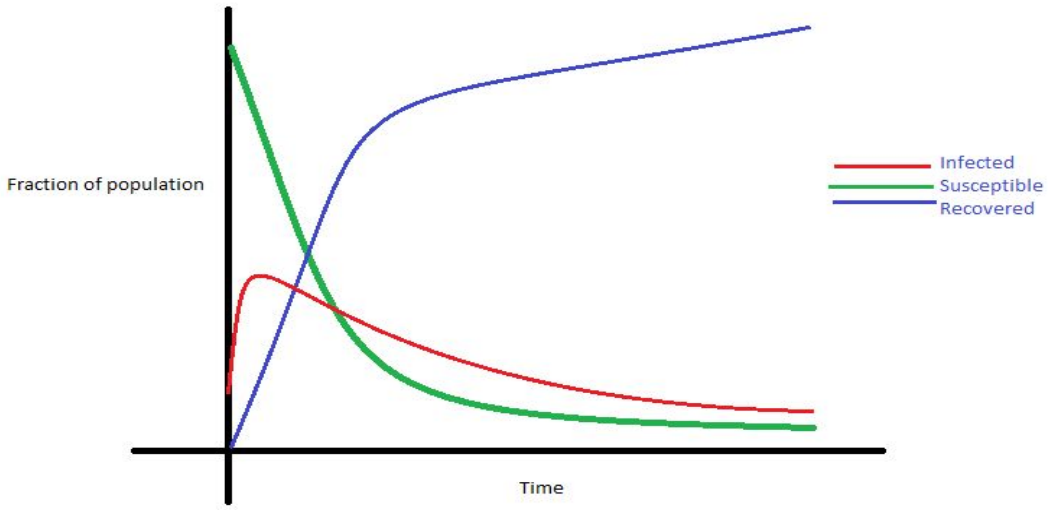
$$1 - R(\infty) = e^{\frac{-\beta}{\gamma} R(\infty)}$$

or

$$R(\infty) = -\frac{1}{R_0} ln(1 - R(\infty))$$

This was the equation that Kermack and McKendrick derived in their theory with specific possible assumptions. They showed that an epidemic with a given R_0 will infect a fixed fraction $R(\infty)$ of the susceptible population by solving the above equation.

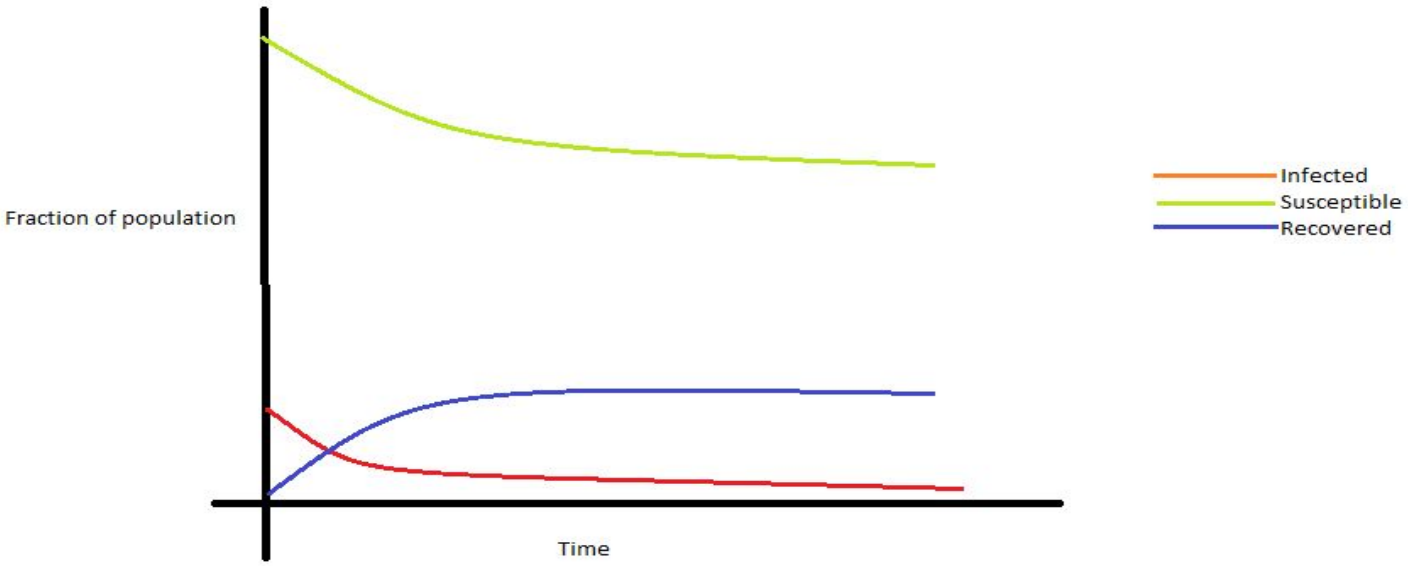
According to the above equation, final outbreak size tends to zero if $R_0 \leq 1$ and increases exponentially when $R_0 > 1$. Therefore, a larger R_0 leads to a larger outbreak which infects the entire population.
Unfortunately, the equation relating R_0 to the final outbreak size from Kermack and McKendrick is only valid when certain assumptions are always true which is not always the case.



For the above graph:

- $\frac{\beta}{\gamma} = 4$
- $i_0 = 0.1$

The above graph clearly depicts that when $R_0 = \frac{\beta}{\gamma} > 1,$ The epidemic spreads through the entire community and eventually converts to a pandemic.



For the above graph:

- $\frac{\beta}{\gamma}$ =0.5
- $i_0 = 0.1$

The above graph unlike the previous one clearly depicts that when $R_0 = \frac{\beta}{\gamma} < 1,$ The epidemic fails to spread over the entire susceptible population and dies out stochastically.

1.2 Motivation of Research Work

The study of epidemiology has created vast scope of research works in network theory more than anything. Scientists spend a lot of time creating vaccines for the epidemic and to do that successfully they must be aware of the network epidemiology which in turn requires a good deal of knowledge of graph theory. We use graph networks to represent a susceptible , infected and recovered population model.
So , network theory and epidemiology go hand in hand.

The network structure of epidemics defines potential transmission routes of infectious disease and that is the reason knowledge of its structure can be used as part of disease control measures.

Contact tracing can be therefore another area of research work as it is a highly effective public health measure. It identifies transmission connections from known infected people and thereby tracing those who are likely to be infected by those who are known infected and hence reducing the spread of epidemic.

1.3 Organisation of the Report

- Kermack and McKendrick Theory for estimating outbreak size
- Epidemiological Analysis from network theory
- Analysis Of Cumulants And Derivation Of Kermack-McKendrick
- Localisation

Chapter 2

Project Work

2.1 Kermack and McKendrick Theory

In the 1920s, The two scientists Kermack and McKendrick did remarkable work in epidemiological studies. They derived a formula of how we can find out the final size $R(\infty)$ of an epidemic with $R_0 > 1$.

They assumed certain things while derivation. These assumptions are as follows:

1. The disease is transmitted in a population with no birth or migration .i.e, the population must be closed.
2. The population is large enough to justify a deterministic analysis.
3. The disease results in complete immunity or death,
4. All individuals are equally susceptible to disease and equally immune to disease.

However ,it was noticed that simply relying on R_0 is quite misleading because even when R_0 was held constant, if there were even a significant change in the assumptions of the theory , the results had a vastly different outbreak size.
This is where network epidemiology comes into picture which determines that R_0 is not the only factor on which the final outbreak size depends . Heterogeneity in the number of secondary infections is also responsible for vast outbreak sizes of the pandemic. To more fully quantify how heterogeneity in the number of secondary infections affects outbreak size, we turn towards network epidemiology and derive an equation for the total number of infected individuals using all moments of the distribution of secondary infections. However , R_0 was only the first moment which cannot be completely used to determine outbreak sizes.

2.2 Epidemiological analysis from network theory

To study epidemiology from a network theory point of view, we must define probability generating function of the degree distribution.
Probability generating function is an alternative representation of a prob-ability distribution. Take the distribution of vertex degrees in a graph. The Corresponding generating function is:

$$G_0(x) = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + \ldots \; p_k x^k$$

Where p_k is the degree distribution i.e, probability of a node having k contacts. So, $G_0(x)$ can also be called as the distribution of vertex degrees in the graph.
The PGF can be used to generate all the probabilities of the distribution and it tells us everything there is to know about the distribution.

Moment generating functions

- Base: $G_0(1) = \Sigma P_k = 1$ (It is the sum of probabilities).
- First moment, $< k > = \Sigma k\, p_k =$ $G_0`(1)$
- The n-th moment $\langle k^n \rangle = \Sigma k^n\, p_k = (x\frac{d}{dx})^n\, G_0(x)\,|_{x=1}$

If we consider a node of degree k, then following any random edge will lead to a node of degree k with probability which is k times than that of reaching node of degree 1 following the same edge.
So, It is k times more likely to follow the edge to a node of degree k than a node of degree 1.

- Probability that a random edge is attached to node of degree k is $\frac{k\, p_k}{\Sigma k\, p_k}$.
- There are k-1 other edges outgoing from this node called excess degree.

Here comes another degree distribution called excess degree distribution $G_1(x)$.
The number of edges around a particular node other than that edge through which we arrived at that node is called excess edges.
The excess degree is 1 less than the actual degree
Let m_k denote the following a random edge that will lead to a node of excess degree k.

So, $m_k = \frac{(k+1)\,p_{k+1}}{<k>}$

So, $G_1(x) = \sum_{k=0}^{\infty} \frac{(k+1)\,p_{k+1}}{<k>} x^k$

$= \sum_{k=1}^{\infty} \frac{k\,p_k}{<k>} x^{k-1}$

$= \frac{1}{<k>} G`_0(x)$

Thus, $G_1(x)$ is the generating function that is responsible for the number of secondary infections from primary cases of infection.

2.3 Analysis Of Cumulants And Derivation of Kermack-McKendrick

The network model for epidemiology naturally removes all the assumptions of kermack and McKendrick while estimating the relationship between R_0 and $R(\infty)$.
However, $R(\infty)$ is not only the function of R_0 but also it depends upon the heterogeneity in the number of secondary infections. Heterogeneity in the number of secondary infections means that some individuals are more likely to spread diseases quicker if infected than others.
The network approach also accounts for stochasticity in the model. It means that if R_0 >1 , then the epidemic will infect the entire population and will convert into a pandemic. There is still some chance that the epidemic will die out even when R_0 >1. This is possible when patient zero who started the infection is outside of the giant connected component and only leads to small outbreak size.
The classical model in terms of cumulant generating function is more useful for epidemiological studies as compared to Probability Generating Function model.

The Cumulant generating function for a random variable X can be written as

K(y) = $\frac{\Sigma k_n\, y^n}{n!}$

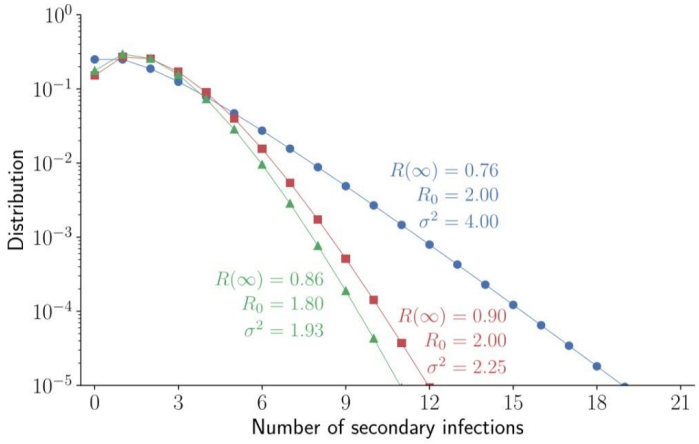
where k_n are the cumulants of the distribution of secondary infections.

$k_1 = R_0$

$k_2 = \sigma^2$

k_3 = skewness of the distribution

k_4 = kurtosis



The above graph works well in accordance with both the cumulant generating function model of epidemic as well as the Kermack-McKendrick model. It depicts that the final outbreak size does not only depend upon the basic reproductive number but also is some function of heterogeneity in the number of secondary infection(σ^2) and significantly on the other cumulants also. There is no direct relationship between R_0 and $R(\infty)$.

A Probability generating function is linked to Cumulant generating function by a direct relationship.

$$G(x) = e^{K\ ln(x)}$$

Therefore,we can replace $G_1(x)$ for the distribution of secondary infections by a function in terms of the cumulants distribution function .

So , in terms of cumulants

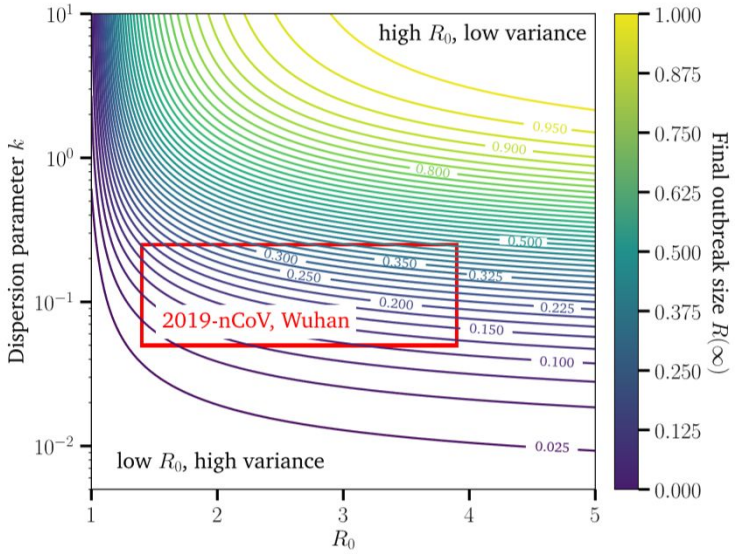
$$G_1(x) = e^{\Sigma \frac{1}{n!} k_n (ln(x))^n}$$

$$= \exp\left[R_0 |\ln x| - \frac{1}{2}\sigma^2 |\ln x|^2 + \frac{1}{6}\kappa_3 |\ln x|^3 - \frac{1}{24}\kappa_4 |\ln x|^4 \ldots\right]$$

The moments of the above equation can be interpreted as follows.

- For a disease to spread , merely R_0 is not the only important factor
- A disease needs a high average number of secondary infections to spread but not always. It depends on other cumulants of the distribution.
- For a fixed R_0 , a disease with small variance σ^2 in secondary infections will spread much more rapidly.
- If variance of heterogeneity in the number of secondary infections is given to us ,then the disease with high skewness will be more stable than a disease with negative skewness
- Given a skewness, a disease will be less stable if it has frequent large positive deviations rather than infrequent small deviations, i.e, a smaller value of kurtosis

Hence, the above results and interpretations show that even if the R_0 value is low for a particular epidemic, other higher moments of the distribution of secondary cases ,easily invade the population and to reach a larger final outbreak size.



The above graph depicts the final size of outbreaks $R(\infty)$ with different R_0 and heterogeneity in the distributions of secondary cases. It clearly summarises all our results and interpretations.

2.4 Localisation:

Relying on R_0 to describe the epidemic and its consequences means that we consider the population to be an average population or rather , every individual is equivalent to an average individual. This means we are considering all the contacts also to be equivalent and equally likely. This is called ***mass-action approximation***. This term means that for any emerging epidemic or a sudden outbreak, we consider the population to be a mixed population ignoring its complex structures ,social interactions. So, here we are ignoring the underlying heterogeneity while considering the average of a distribution. As a result, many mathematical issues will arise which we need to tackle through a still better epidemiological model which is comparatively difficult to design as well as implement.

Network science provides a framework that removes the defectiveness caused due to mass action approximation introducing a new feature called ***heterogeneous pair approximation.*** This considers the structure of contacts among individuals. So, we began characterizing individuals on the basis of their number of contacts.

Unlike the classic model of epidemics which only uses the data of fraction of susceptible S and infectious I individuals to carry out the result of determining outbreak size, in the new model we can now track the proportion of susceptible and infectious individuals with k number of contacts (S_k and I_k respectively) and the fraction of contacts in the network connecting

- Two susceptible individuals ([SS]) or
- Two infectious individuals ([II])
- One of each ([SI]).

As an outcome of the standard epidemiological model which shows there is a straightforward relationship between basic reproductive number and final outbreak size, there exists a critical value β_c for the transmission rate below which epidemics are bound to localise in a certain substructure of a network.
Also it is obvious that larger structures having larger numbers of nodes n are likely to contain more infectious individuals than smaller one.

Chapter 3

Conclusions and Discussion

There are over 110 different infectious diseases in 2019 itself. So, there is a rapid increase in the need of health management. That is why the term R_0 is used most effectively in public health. However, in this project we come know that R_0 is not just ineffective in determining the size of the outbreak alone but it is more probable to have larger sized outbreak with low values of R_0 considering the heterogeneity in the number of secondary infections. We used network science to derive the probability of an epidemic converting to pandemic and determining its final size.

Contact tracing can be an effective tool to handle the epidemic to a great extent. It is a preventive measure of epidemic control. It means isolating everyone whom we know that they have been in contact with an infected person. Thereby decreasing the chance of spreading the disease to more people.

The need of contact tracing directly informs us about potential secondary cases caused by a single individual even before they become serious, and therefore provides us with an estimate for $G_1(x)$. Both for generating formal predictions of epidemic risk and controlling the size of the outbreak ,it is very important for us to begin contact tracing before numerous transmission chains become widely distributed across the world. We can also use past data related to similar outbreak as that of nCoV-2019 to predict the epidemic size. Using past data we can say that the range of the final outbreak size must be between 5% to 50% of the total susceptible population. This large range is in accountability of heterogeneity in the number of secondary infections. However, predicting outbreak size based on early data is an incredibly complex task and cumbersome at the same time but also it is within reach of most of the scientists of the world due to new development of mathematical analyses and faster communication of public health data.

Bibliography:

[1] Danon L, Ford AP, House T, et al. Networks and the epidemiology of infectious disease. *Interdiscip Perspect Infect Dis*. 2011;2011:284909. doi:10.1155/2011/284909

[2] St-Onge, G., Thibeault, V., Allard, A., Dubé, L. J., & Hébert-Dufresne, L. (2020). Master equation analysis of mesoscopic localization in contagion dynamics on higher-order networks. *arXiv preprint arXiv:2004.10203*.

[3] Derivation of excess degree distribution

a. ECS 253 / MAE 253, Lecture 15 I. Probability generating function recap