UNIVERSITY OF DHAKA

A Risk Score Based Model for Dynamics of Infectious Disease Spread

by

Exam roll: CURZON 102717
Registration No: 2017-314-985
Session: 2017-18

Exam roll: CURZON 102702
Registration No: 2017-015-022
Session: 2017-18

A thesis/report submitted for the degree of Bachelor of Science at University of Dhaka

Department of Computer Science and Engineering

Faculty of Engineering

Declaration Form

University of Dhaka

We, Fariha Anjum and Humayra Jahan , declare that the work given in this report is the result of an investigation conducted by us in the year 2022 under the supervision of Dr. Upama Kabir . We further affirm that no component of this work has been or will be submitted for a degree or certificate elsewhere.

Fariha Anjum
Registration No: 2017-314-985
Candidate 1

Humayra Jahan
Registration No: 2017-015-022
Candidate 2

Dr. Upama Kabir
Professor
Department of Computer Science and Engineering
University of Dhaka
Supervisor

UNIVERSITY OF DHAKA

Abstract

Infectious viral diseases that are newly developing or returning have threatened humanity throughout history. Infectious diseases significantly impact the economics and quality of life of a nation. The threat of infectious diseases has increased considerably due to a multitude of factors. The infection rate can be significantly reduced if it is addressed beforehand. Human movement and infectious diseases are strongly intertwined. Human migration can influence the dynamics of infectious diseases by introducing infections into vulnerable groups or increasing the frequency of interactions between infected and susceptible people. Researchers have used this information to battle infectious diseases. Recent studies have focused on post-pandemic situations. However, early detection of links across different zones can help with disease prevention. This research exercises the exposure level of different nearby areas of a specific region to get insight into the progression of infectious diseases. We introduced a method to compute risk score considering the risk probability to each node in an area. This risk score assesses the level of risk in a particular zone due to an outbreak. The score indicates the vulnerability level of a zone. The performance analysis reveals that the suggested model performs well on large data sets and significantly contributes to the early detection of an outbreak's spreading path.

Acknowledgements

We would like to express our deep and sincere gratitude to our supervisor, Dr. Upama Kabir for providing invaluable guidance and continuous support. We would also like to give very special thanks to Dr. Mosarrat Jahan for her encouragement and insightful comments.

We would also like to thank our Samin yaser and Farhan Fuad Haque who are the students of Department of Computer Science and Engineering of University of Dhaka for all their support in completing this project.

Finally, we would like to thank our parents for supporting us spiritually throughout our life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Throughout history, new and reemerging infectious viral diseases have threatened humankind. Disorders that are caused by pathogenic microorganisms are called contagious diseases. Bacteria, viruses, fungi, and worms/ helminths are the most common types that cause infectious diseases. Some diseases are transmitted from person to person; some are from animals or insects to humans. [18]. Even people can infect diseases by consuming contaminated foods or any injury exposure to the environment. Major pandemics and epidemics such as plague, cholera, flu, ebola, severe acute respiratory syndrome coronavirus (SARS-CoV), and Middle East respiratory syndrome coronavirus (MERS-CoV) have already afflicted humanity [11]. These diseases have reminded us of the importance of remaining watchful against introducing new infectious diseases. Various strategies have been developed to control the spread of a viral sickness. Quickly identifying, isolating, and treating cases and providing supported quarantine of contacts break the chains of transmission, effectively stopping virus transmission.

Major pandemics and epidemics have previously impacted mankind, including the plague, cholera, flu, severe acute respiratory syndrome coronavirus (SARS-CoV), and Middle East respiratory syndrome coronavirus (MERS-CoV) [19]. The new

coronavirus illness of the 2019 (COVID-19) pandemic is currently affecting the whole planet. Due to increased human interaction with animals due to breeding, hunting, and international trade activities, zoonotic pathogens, which are infectious illnesses that can spread during pandemics, have been transferred to humans. The development of infection prevention and management techniques was made possible by our understanding of the viruses' transmission processes to people. Implementing public health measures like isolation, quarantine, and border control over the years has helped to keep contagious illnesses under control and preserve the social order. These containment techniques are still used today to manage the COVID-19 pandemic without pharmacological therapies. To quickly identify the onset of infectious risks, global monitoring programs of water-borne pathogens, vector-borne illnesses, and zoonotic spillovers at the animal-human interface are crucial. For a successful response in the event of pandemics, new technologies for quick diagnostic testing, contact tracing, medication repurposing, biomarkers of disease severity, and new platforms for developing and manufacturing vaccines are required.

Philip S Brachman discusses contagious diseases at [2] that have significantly impacted population dynamics throughout human history. Hippocrates recognized a connection between climate, diet, and living conditions in his writings about how diseases spread through the air, water, and places. In discussing germ theory in the 1500s, Fracastoro hypothesized three transmission pathways: direct touch, fomites, and transmission at a distance (airborne). Epidemics of cholera, smallpox, syphilis, yellow fever, typhoid fever, leprosy, plague, and other infectious diseases were common. Chemotherapy and antibiotics were introduced to our arsenal against contagious diseases during the 20th century. We now understand that re-emerging and emerging infections have become a major global hazard. Recently [8], the development of disease-fighting strategies has received significant interest from the research community, virologists, and medical professionals due to COVID-19. The pandemic compelled the breadwinners to stay indoors and put a stop to commerce and business. The ecosystem recovered as a result of the lockdown.

Technology [14] refers to the methodologies, systems, and equipment that result from scientific information being applied to practical goals. Machine learning (ML),

natural language processing (NLP), and computer vision applications are examples of artificial intelligence. These capabilities teach computers how to construct, represent, and anticipate using large information-based models. Additionally, it talks about using the internet to alert, raise awareness, and govern social behavior. Here are a few ways that technology is being used to combat COVID-19.

1.2 Motivations

Vaccine researchers, regulators, and health departments face unprecedented hurdles as infectious diseases spread at alarming rates. In the last 20 years, the number of outbreaks and pandemics caused by emerging and re-emerging diseases has increased dramatically, causing tremendous economic and societal damage. As a result, projecting the rate and trajectory of an epidemic's infection rate plays a vital role in managing an epidemic. The most effective approach to guard against a contagious disease is to be informed of its transmission before and take appropriate precautions to combat it. The government and the general people can take necessary precautions if they can be told of the outbreak sooner. It will be easier to handle the outbreak and pandemic situation if it is known the next possible infected area after any new emerging or re-emerging disease detected in any area. This necessitates the development of a system that can forecast the spread of a dangerous disease.

1.3 Problem formulation

Infectious diseases substantially impact a country's economy and residents' lives. Due to various causes, the danger of infectious illnesses has dramatically risen in recent years. For example, the recent outbreak of emerging infectious disease, i.e., severe acute respiratory syndrome coronavirus 2 (later named coronavirus disease 2019, COVID-19), has brought significant challenges to the global economy and

public health. Without protective immunity at both the individual and population levels, an emerging infectious disease may spread rapidly, resulting in many severe cases and deaths in the community [1]. Predicting infectious diseases can effectively control large-scale outbreaks and reduce the transmission of epidemics in rapid response to serious public health events. Therefore, experts and scholars are increasingly concerned with predicting infectious diseases [26] . It is critical to understand how the disease will spread and which areas are more vulnerable than others. This would be hugely valuable if the authorities could predict the rate at which the disease would spread from an affected area to neighboring locations. Therefore, it is imperative to identify cases with a risk of disease progression for the optimized allocation of medical resources in case medical facilities are overwhelmed with a flood of patients.[7]

1.4 Research Objectives

We understand that by identifying the pattern of infection, the propagation of contagious diseases can be prevented. Novel virus outbreaks are linked to humans due to demographic density, travel, commerce, and external factors such as environmental issues and farming practices. That is why emerging disease surveillance is crucial for the early diagnosis of public health hazards. Forecasting infectious diseases is a field that will never be obsolete. In recent years, various approaches for preventing infectious diseases have been proposed, and we want to take it a step further.
In this project, we develop a model to calculate the risk level of the neighboring areas of any infected area based on mobility data. Our objectives for this project are,

- To help the government by giving alerts before any outbreak happens.
- To reduce the damages of any pandemics for a new infectious disease.
- To help the citizens to alert about the spreading of the disease.
- To develop a model to give a fast solution to prevent disease spreading.

1.5 Contributions

Infectious illnesses that could become pandemics have frequently emerged and spread throughout history. New diseases are discovered from time to time. So it concerns humankind. People cannot stop unknown diseases with our recent technology, but we can control them after discovering any disease. So in our study, we propose a technique to control the spreading of new emerging or re-emerging diseases. In our work, we aim to give an amiable and fast solution to tackle the outbreak before it happens for a new contagious disease. So that the spreading of the disease and damages can be minimized by alerting the government and citizens. To achieve that goal, we proposed a model to calculate the risk level for every area before any outbreak happens. It will help to predict the next possible infected area after detecting infectious diseases in any area. It will help the government and the citizens of a country to be alert on the early stage of disease detection in any area before it spreads to a new area. Thus, the government can take necessary steps to prevent the disease from spreading and causing economic and social damage.

1.6 Organization

The remaining sections of the project report are structured as follows. Relevant research papers and background studies are described in Chapter 2. Chapter 3 presents our suggested system's system concept, issue formulation, and solution features. Chapter 4 presents the performance of the suggested approach, along with comparisons with different threshold values. Chapter 5 summarizes our efforts and provides an overview of our future work goals.

Chapter 2

Literature review and background study

In recent years, several illness detection technologies have been developed. We briefly overview the current studies' methodologies, contributions, challenges, and limitations.

2.1 Risk Prediction and Outbreak Handling Methods

M Woolhouse at [25] discussed about some possible ways to predict the risk of contagious disease. Various methods have been applied in practice to forecast illness risks in the future. These consist of statistical techniques, risk modeling, simulation modeling, and expert opinion. The types of information needed to forecast disease risks can be separated into three categories: that which relates to the disease, the host, and that which relates to the environment. The emergence of novel pathogens is an important, current, and difficult application of techniques to anticipate disease risk. Alarming rates of pathogen emergence and reemergence are occurring. Numerous research on certain pathogens have connected such pathogens' emergence or re-emergence to distinct factors. In summation, formal, quantitative methods

have gotten more complex and well-liked in recent years when used to forecast the dangers of infectious diseases. Although epidemiologists are the experts in this field, many other fields are greatly influenced by it.

E Rees at el [20] discussed about early identification and forecasting of infectious disease epidemics. Increasing the diversity of data utilized in modeling methodologies is a significant achievement for risk assessment. The process of identifying and defining characteristics in people or communities that enhance their vulnerability to catching disease is known as risk modelling in the context of infectious diseases. Regression analysis is a component of the well-founded and illuminating risk modeling strategy known as statistical inference. This technique is employed to ascertain the relationship between risk variables and the desired outcome. Information from open-source internet data is increasingly being used into regression models and statistical inference in general. An early example was the use of Google Flu Trends search query data as a prediction of the number of reported doctor visits for flu-like symptoms. Using compartmental models to quantitatively replicate population transmission dynamics—that is, the movement of people between health stages like susceptible, infectious, and recovered—is another popular risk modeling strategy. Public health experts require greater access to current surveillance data in order to monitor infectious disease outbreaks in an efficient and prompt manner.

V. J. Brookes at el [3] suggested incorporating these tools into a framework that improves the creation of tactical and strategic plans for preparing against new risks of emerging and re-emerging infectious disease (EID). Information is a vital prerequisite to quickly detect the presence of EIDs and foresee potential future dangers to the health of people, animals, and the environment. Developing tactical and strategic plans that are appropriate to the social, cultural, economic, and environmental context in which prevention and control activities take place requires an understanding of the significance of the range of potential impacts of emerging threats and EIDs on those affected. The last several decades have seen an increase in the use of risk analysis techniques in the fields of animal and public health to determine

the likelihood of an unwanted event occurring, its possible wide-ranging effects, and feasible mitigation measures. These techniques offer reproducible, transparent, and impartial evaluations.

In [21] Sripanidkulchai, K The danger of healthcare staff coming into touch with COVID-19 patients is very high. They sought to uncover independent parameters connected to COVID-19 infection in hospital personnel following occupational exposure(s) because there is currently no evidence-based, comprehensive risk assessment tool for exposure related to healthcare. They also sought to develop a risk prediction model. High levels of education and potent immunizations provided infection protection. They developed a risk model and scoring system with strong discriminating capabilities. Following exposure events associated to healthcare, there was a higher risk of laboratory-confirmed COVID-19 because to symptoms, unprotected exposure, lesser education, and obtaining low strength vaccinations. The goals of this study are to uncover independent risk variables and to develop a quantitative risk model and risk score for healthcare-related exposure. SARS-CoV-2 infection was diagnosed through RT-PCR in hospital staff after exposure to confirmed positive patients. In this work they calculated risk score for person to person.

2.2 Community Detection and LPA

Htwe Nu Win and Khin Thidar Lynn at [24] discussed about community detection of facebook. Here, social network is represented as a graph. The facebook friendship graph is unweighted and undirected graph. The groups of friendship are considered as communities in social networks. Nodes that cannot group with any communities and are not required for identifying the communities in graphs they will be recognized as outliers within the group. The use of neighborhood overlapping for identifying outliers and noteworthy communities is explored in this research utilizing the vertex similarity approach. The majority of outlier identification algorithms

use Hawkin's definition of outliers, and our system bases its assumptions on the idea that outliers are nodes with 0 values for the similarity measure in the graph. Existing distance or density-based methods can quickly identify these outliers. In this article, loneliness users are modeled as solitary nodes. Since they have no nearby neighbors, their edge structure cannot be taken into account. Communities are thought to be collections of vertices that are related to one another. After looking through seed nodes, they calculated the similarity between each pair of vertices. The majority of currently used similarity methods are based on distance measurements like the Euclidean, Manhattan, and others. The notion of an undirected graph was employed in this study, where two vertices have a common node if they share one or more of the same vertices. They primarily employed two techniques: first, vertex centrality, which is used to identify seed nodes with the greatest degree or number of neighborhoods, and second, neighborhood overlap based on vertex similarity. To find the efficient community, they suggested an edge structure-based outlier detection approach. The weakness of this research is that it relies on the edge structure of an unweighted, undirected graph to identify notable communities and outliers without considering the topology. Their technique has the flaw of being dependent solely on edge structure without taking into account any information, such as the user's profile.

Andreas Kanavos at [13] compares six well-known community discovery techniques, including Breadth-First Search, CNM, Louvain, MaxToMin, Newman-Girvan, and Propinquity Dynamics, using four popular graphs and data gathered from Twitter concerning both artificial and natural data. When it came to user-based evaluation, they showed some students the communities that each algorithm had extracted, along with a corresponding user and their tweets in the grouping and took into consideration three different alternatives for the extracted communities: "dense community," "sparse community," and "in-between." According to their research, community-detection algorithms can help in locating dense user groups. Network centrality, which indicates independence, autonomy, dominance, and influence in a network, can be used to evaluate the power of the network based on the relationships between each node. One of the often used indicators in connection to

network data analysis is centrality metrics. They point out the necessity of parameterize some factors, such as status, visibility, structural strength, or prestige, by emphasizing the unit's predominance as a major factor in centrality analysis. The researchers of this paper evaluated five widely used community-detection methods based on higher-order information that is found as graph constructions. Clement Newman Moore's suggested algorithm (CNM) provides a way for dividing vertices into distinct communities for each one. This algorithm permits examination at several levels, from "local" to "international," until it is bound by the criterion. The Louvain method, also known as the multilevel algorithm, uses weighted graphs for its hypervisor-based grouping analysis. High peripheral densities are desired for neighborhoods, although there are currently few dense areas within communities. The MaxToMin method attempts to link a community to the nodes with the most powerful surrounding edges. The algorithm may travel the whole length of the graph using this method. In practice, it moves from the margins of the strongest to the weakest, but it is unable to move in the opposite direction. The Newman-Girvan (NG) algorithm, also known as the edge betweeness algorithm, is based on betweenness centrality. If the graph in this algorithm is broken, the process described above must be repeated for each linked component. The PD algorithm is used in community discovery approach by calculating the probability that two vertices are a cohesive community. Without assuming any knowledge of a community's layout, this approach absorbs similarity information from the network topology in a spontaneous manner. The Newman-Girvan and Propinquity Dynamics approaches are demonstrated in this study to be confirmed and proven to generate optimal performance on practically all datasets. Additionally, the lowest values are displayed by the CNM and Breadth-First Search techniques. MaxToMin with Breadth-First Search has the lowest values while Propinquity Dynamics and Newman-Girvan have the greatest values in the dolphins graph, respectively. In order to identify the variables that impact the findings of the paradigms at a more granular level of detail, the authors intended to conduct a substantial collection of additional experiments under different settings (thematics). The recommended project can benefit greatly from the use of effective heuristics to time-varying graphs. Research can incorporate

experimental, analytical results from the theory of dynamic systems, or even other analytical algorithmic methods.

Fei Wang and Changshui Zhang at [23] proposed a modified version of label propagation algorithm (LPA) as Linear Neighbourhood Propagation (LNP). In order to solve the edge weights in each patch, a typical quadratic programming technique is used. The LNP approach first approximates the entire graph using a series of overlapped linear neighborhood patches. After that, the weight matrix for the entire graph will be created by averaging all of the edge weights. It is intriguing to learn that LNP technique still achieves good identification accuracy even when it is simply labeled a very small portion of the data. Only a small number of labeled points are necessary to accurately estimate the labels of the remaining unlabeled points thanks to LNP's ability to successfully reveal these complex structures.

Yasuhiro Fujiwara, Go Irie at [10] approached an effective LPA in order to determine a label for each node, lower and higher bounding scores are computed. This iteratively prunes unneeded score computations. The labeling scores are repeatedly calculated using the power approach until convergence for all labels. They didn't update the scores for all labels to increase efficiency. Instead, they updated a selection of labels' labeling scores using new method. The lower and upper boundaries of labeling scores are used to establish subset membership. In contrast to the power technique, it stopped iterations early if there was no label that needed to be modified. This suggests that fewer repetitions are required for this strategy than for the power method. Second, it achieved the exact same labeling outcomes as the ideal solution. This is so that unneeded score computations might be securely discarded thanks to the lower and higher limitations. Finally, no user-defined inner parameter is needed with this technique. The power technique, in contrast, necessitates the selection of a specified threshold for iteration termination, which results in a trade-off between accuracy and efficiency. In other words, this strategy is user-friendly.

D. Malhotra and A. Chug at [17] proposed a model of modified label propagation for community detection. To identify communities in complicated networks, they combined structural characteristics with node properties. In order to generate a

weighted graph, the algorithm first translates the textual data. To solve the random selection issue, it then applies the modified LPA with various connection strength metrics. The method then provides the community architecture of the network with the specified node attributes. Further research can concentrate on expanding the current paradigm to detect communities that are overlapping whereas every node might link itself with many communities, similar in real-world systems.

Jung et al. [12] presented a cluster-based analysis system of infectious disease occurrences that can uncover commonalities or differences between clusters by grouping items with comparable occurrence patterns.They collected and preprocessed data on infectious illness occurrences based on time, geography, and disease. Then, using Tucker decomposition,they extracted latent variables in the dimensions of time, geography, and disease. They run k-means clustering on these latent features and assess the results for each dimension. Some sickness clusters were seasonal and recurring, whilst others were aperiodic.

Vince Lyzinski, Gregory Sell, Aren Jansen at [16] many cutting-edge graph clustering techniques for unsupervised term discovery systems are thoroughly evaluated. Scalability of algorithms is crucial because even tiny corpora can produce graphs with tens of millions of edges and millions of millions of nodes. In addition, since manual validation is required for real-world applications, the unsupervised nature of the problem favors algorithms with fewer tuning parameters. In this study first they constructed a graph and then applied modularity based clustersing and label propagation algorithm (LPA). Modularity is a commonly used criterion to assess how well graph clustering methods work. Heuristically, the proportion of edge weights that fall within clusters as opposed to across clusters is used to determine a clustering's modularity. For non modularity clustering they used weighted variant of LPA. In second method, firstly algorithm created an initial state where each vertex is assigned to a distinct cluster. Each node updates its cluster assignment by maximising over cluster labels c, iteratively and sequentially over all vertices. This cluster assignment updates asynchronously in this method and with few repetitions,

good performance is frequently achieved.

Chuanwei Li at el [15] suggested a technique for stable community detection that combined label propagation with density peaks (DS-LPA). A quick and effective community detection approach without parameters is the traditional LPA. The approach converges by first propagating node information depending on the network topology and then using an asynchronous update strategy. On the other hand, the density peak clustering algorithm has two crucial components: the local density of data points and the minimal distance (DPC). In DS-LPA algoritm, firstly to choose the community center, the local density and minimum distance of each node are computed. Then, an initial community is created in accordance with the community center, and finally, the remaining nodes are informed about the initial community's label. This paper suggests a reliable label propagation method to get around LPA's drawbacks caused by high unpredictability. First, a foundational community formed around the chosen community center is created. The update order is then established using the node's propagation power. Thus the LPA label is completed updating rules by adding this.

Chapter 3

Proposed Approach

3.1 Introduction

In this chapter, we present the methodology of our proposed scheme.the above mentioned problem.

3.2 Proposed Method

The infection rate can be regulated to a large extent with the aid of a tool that identifies the interconnected zones of a specific location. Infectious diseases and human mobility are intimately associated. This information has been used by researchers to combat infectious diseases. By bringing infections into susceptible groups or altering the frequency of encounters between infected and susceptible people, human movement can influence the dynamics of infectious diseases. To understand the connection between the spread of infectious diseases and human mobility, we considered the statistics on mobility. Recent studies focus on post-pandemic conditions. However, the early finding of linkage between various zones can aid in disease control. In order to gain insight into the progression of infectious disease, our objective is to identify risk level of different adjacent areas of a certain area.

3.3 System Model

3.3.1 Parameter selection

Human mobility is important in the temporal and spatial spread of infectious diseases. Over the last few decades, researchers have extensively researched how human mobility affects disease transmission[6]. Increased global mobility of people, animals, plants, and commerce is causing infectious diseases to spread to new areas. Because of mobility, infectious illnesses have recently moved well beyond the previously recognized geographic bounds. Travel and population movement, in particular, are now essential factors in spreading bacterial infection [9]. Understanding the roles of human transportation in the transmission of infection is essential for creating effective disease control schemes[6]. Given the importance of mobility in spreading infection, we chose the mobility data to population ratio as an essential feature in our approach.

3.3.2 Dataset

The proposed system focuses on mobility data between two areas to find the risk probability. We had to collect mobility data between two areas and the population of each area. The population is collected from the 2011 census [5]. And the mobility data is gathered from the online bus ticket booking website busbd.com [4] . We gathered information from a specific day to get a sample of the number of people traveling from one area to another. We have 348 areas and 3318 connections between them. We represent this in a graph of nodes and vertices.

3.3.3 Initialization

We have a graph G= (V, E)with a set of vertices V= v1, v2,v3...vn and edges E= e1, e2, ... en. Each edge contains an ordered pair of connected vertices. Social networks can be treated as graphs in the real world, with people as nodes and their

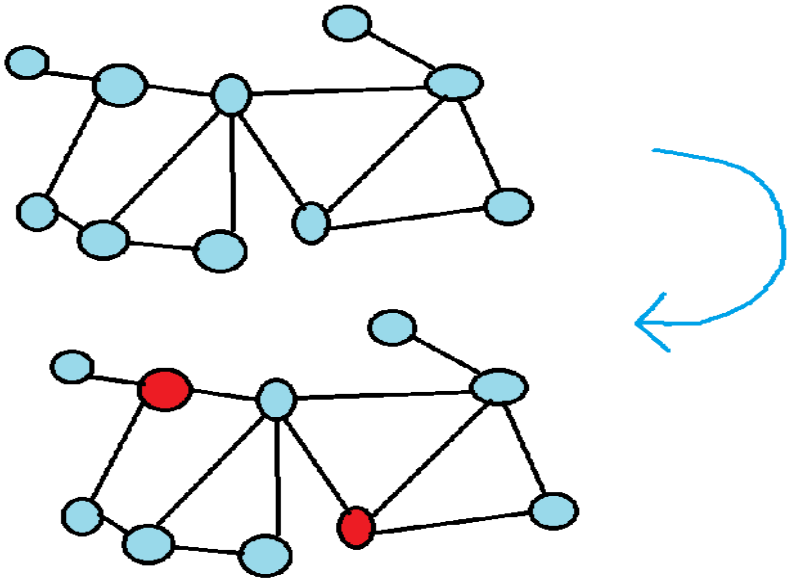


FIGURE 3.1: Infected nodes set as stable nodes

interactions as edges. So, in our problem, we treat each area as a vertex of the graph and the connection between each area as an edge. Infected vertices are chosen as stable nodes. Other nodes are chosen as unstable and gradually checked as stable based on a similarity threshold.

3.3.4 Risk Score calculation for all vertices

3.3.4.1 Risk Score ξ

In response to significant public health crises, infectious disease prediction can successfully control large-scale outbreaks and reduce epidemic transmission. As a result, scientists and scholars are becoming increasingly interested in infectious disease prediction [26]. Understanding how the disease will spread and whether locations are more vulnerable than others is crucial. This would be extremely useful if authorities could forecast the rate at which the disease spreads from an affected area to nearby areas. Algorithm 1 measures risk for each area based on their mobility data.

$0 \leq \xi \leq 1$

Infected nodes are labeled 1, and uninfected nodes are labeled 0. A score closer to zero indicates that the area is more susceptible to infection than other places.

3.3.4.2 Weighted mean of risk scores

The weighted mean is a type of mean calculated by multiplying the weight (or probability) associated with a particular event or outcome with its associated quantitative outcome and then summing all the products together[22]. Here, we calculate each vertex's risk score using the following formula. For each vertex V , the n adjacent vertices of vertex V are = 1,2,3...n. So, we take the following parameters to update the risk score

δ_i = mobility / population

η_i = current risk score of node

For each vertex V risk score,

$$\xi_V = \frac{\sum_{i=1}^{n} \delta_i \times \eta_i}{\sum_{i=1}^{n} \eta_i} \tag{3.1}$$

where i = 1,2,3 ... n .

3.3.4.3 Risk score propagation

One of the widely used algorithms for finding communities is the Label propagation algorithm.LPA has gained significant attention due to its advantages of linear time complexity. It lacks the need to define the objective function and the number of communities in ahead of time. We employ the concept of LPA to improve the risk score. It allocates a risk level of 0 to 1 to the vertices. A higher score implies greater risk. We use this strategy to propagate the risk score from infected nodes to nearest neighbors. As a corollary, we can anticipate how other regions may be affected over time. Eventually, all the nodes are stable and the procedure ends.

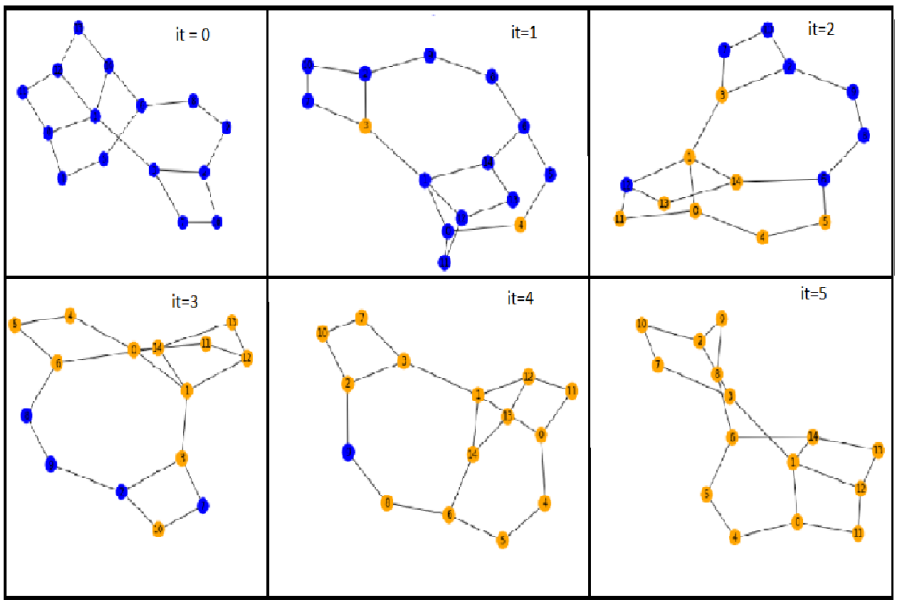


FIGURE 3.2: Risk score propagation

3.3.4.4 Threshold β

After some nodes are infected, the algorithm starts to propagate that infection score to other nodes. And it keeps updating the risk score until all nodes become stable or the change of risk score is very small (β) The risk scores assigned to nodes become stable when the change from the previous value is not greater than threshold β .

3.3.5 Node stability

We select a value of count for setting the node score as stable to define the risk score. This risk score defines the probability of getting an area infected from it's neighboring infected nodes. We experiment the result of the algorithm by using different values of count .

γ, Count of iterations for nodes to be stable

β , Threshold

ξ_{i+1} = new risk score

ξ_i = risk score before updating $\delta_\xi = \xi_{i+1} - \xi_i$

$$f(\beta) = \begin{cases} \text{count} = \text{count} + 1, & \text{if } \delta_\xi < \beta. \\ \text{count} = 0, & \text{otherwise.} \end{cases} \tag{3.2}$$

3.3.6 Algorithm

Using Algorithm 1 we calculate the risk score for each node using weighted mean likelihood in equation 3.1. This is one type of mean calculating formula that performs better than the average mean. We use a modified label propagation algorithm to propagate the risk score of the infected nodes to other adjacent nodes.
We employ the concept of LPA to improve the risk score. It allocates a risk level of 0 to 1 to the vertices. A higher score implies greater risk. We use this strategy to propagate the risk score from infected nodes to nearest neighbors. After all the nodes are done updating, we can get an insight into the infection situation of adjacent areas.

In algorithm 1, we have a set of nodes and edges which represent the data set. Some variables γ , β are initialized with a value. The nodes that are infected already are passed in this algorithm and set a value of 1. We finish our iteration when all the nodes reach stability. For each area, we set the probability/risk score -1 at first. Then for each node, we start to compute the risk score using formula 3.1. For each vertex, there is a count value which is updated using the conditions of 3.2. If the count becomes equal to γ we set that node as stable. We set the threshold based on how much we want the risk values to change before being stable.

Algorithm 1 An algorithm for assigning risk scores

Data: A graph network of geography and a set of infected nodes
Result: Assigned risk score to each node
$infectedNodes \leftarrow stable$
$threshold \leftarrow \beta$
$countStable \leftarrow \gamma$
$initialCount \leftarrow 0$
$stable[] \leftarrow arrayofstablenodes$

while $convergence \neq True$ **do**
  for all nodes in G $\in \{1, \ldots, n\}$
  if $stable[node] \neq true$ **then**
    for all nodes adjacent node in G $\in \{1, \ldots, k\}$
    $riskscore \leftarrow \frac{mobility[node] \times riskScore[node]}{\sum_{i=0}^{k} RiskScore[adjacentnode]}$
    if $abs(prevRiskScore[node] - riskScore[node]) < \beta$ **then**
      $count \leftarrow count + 1$
      if $count == countStable$ **then**
        $stable[node] \leftarrow True$
      end
    else
      $count \leftarrow 0$
    end
    end for
  end
  end for
end

3.4 Conclusion

In this chapter, we formulate our proposed model to calculate the risk probability of different areas of Bangladesh. We have given a probability-based algorithm to find the risk score for areas of Bangladesh in polynomial time. And as we shall see in the next chapter, our proposed model performs very well on the big dataset and produces an efficient result.

Chapter 4

Experimental results and evaluation

4.1 Introduction

In this chapter, we discussed the performance of the suggested system using several metrics. Throughout history, we have seen how infectious diseases can swiftly become an epidemic, making control impossible without early warning and stringent precautions. After specific nodes in a graph representation of a particular area get infected, we assign risk scores to those nodes to intervene early to prevent viral disease spread. In this chapter, we give extensive system evaluation results.

4.2 Risk Score assignment

Equation 4.1 presents the risk score $f(\beta)$, which defines the probability of getting an area infected from its neighboring infected nodes. We experiment the result of algorithm 1 by using different values of count .

$$f(\beta) = \begin{cases} \gamma = \gamma + 1, & \text{if } \delta_\xi < \beta. \\ \gamma = 0, & \text{otherwise.} \end{cases} \tag{4.1}$$

where,

$\gamma \longrightarrow$ Count of calculation for nodes to be stable

$\beta \longrightarrow$ Threshold

$\xi_{i+1} \longrightarrow$ new risk score

$\xi_i \longrightarrow$ risk score before updating

$\delta_\xi = \xi_{i+1} - \xi_i$

4.3 Performance metrics

The performance metrics of our proposed algorithm 1 is :

- **Node stability vs Iterations**: Based on the chosen threshold and γ, the algorithm converges, which implies that all nodes become stable. An essential indicator is the number of locations that reach stability versus iteration. It converges reasonably quickly at a lower threshold and γ. We iterated the procedure for various gamma and beta threshold values and provided the results. According to it, increasing the threshold and γ by time reduces the number of iterations. So, we can attain efficient results with the appropriate combination of variables.

- **Count(γ) vs. iteration**: In algorithm 1, the parameter γ represents the number of times nodes must be stable before becoming permanently stable. Our dataset includes daily mobility statistics from one place to another. We can predict the risk probability of each zone after one day of infection in particular neighboring places with each iteration. As a result, the number of iterations required vs. parameter is an γ essential measure.

4.4 Performance analysis

4.4.1 Dataset

The proposed approach leverages mobility data between two places to calculate risk likelihood. We needed to collect mobility data between two places and demographic statistics for each area. The population is derived from the 2011 census [5] . And the mobility information is obtained from the online bus ticket buying website busbd.com [4]. We gathered data from a single day to sample the number of persons traveling from one location to another. There are 348 locations with 3318 linkages between them. This is represented as a graph of nodes and vertices.

4.4.2 Different values of γ and β

We run algorithm 1 for multiple values of γ and β. and present the results in Table 4.1 , 4.2 and 4.3 . Table 4.1 show the result of experiment with 100 different thresholds β values for $\gamma = 5$.

Similarly, Table 4.2 represents the performance for the value of $\gamma = 10$.

Similarly we did the same experiment for $\gamma = 15$.

Figure 4.1 corresponds to the relation of the number of iterations needed and the selected γ value $= 5$. For the 100 thresholds, we tested our dataset on this algorithm. The threshold range is from 0.00005 to 0.5. Figure 4.1 shows that when the threshold is smaller, the number of iterations reduces very slowly. As the threshold increases, the number of iterations does not change very much. Even though threshold is increased , performance of algorithm 1 does not degrade with smaller threshold. So , this algorithm performs well for smaller threshold . We want to select a smaller threshold because if the change in nodes risk score is bigger ,it will not give us an accurate risk score .That's why we want to choose a smaller threshold .

Figure 4.2 corresponds to the relation of the number of iterations needed and the selected γ value $= 10$. For 100 threshold values, we tested our dataset on this

Number of iterations for γ =5			
Iterations No	Count of node stability	Threshold	Iteration
1	5	5e-05	23
2	5	0.0001	22
3	5	0.00509999	16
4	5	0.0101	16
5	5	0.0151	15
6	5	0.02010000	14
7	5	0.02510000	14
8	5	0.03010000	14
9	5	0.03510000	14
10	5	0.0401000	13
11	5	0.0451	13
12	5	0.0501	13
13	5	0.0550999	13
14	5	0.0600999	13
15	5	0.0650999	12
16	5	0.0701	12
17	5	0.0751	12
18	5	0.0801	12
19	5	0.0851000	12
20	5	0.0901000	12
21	5	0.0951000	12
22	5	0.1001000	12
23	5	0.1051000	12
24	5	0.1101000	12
25	5	0.1151000	12
26	5	0.1201000	12
27	5	0.1251000	12
28	5	0.1301000	12
29	5	0.1351000	12
30	5	0.1401000	12
31	5	0.1451000	12
32	5	0.1501000	12
33	5	0.1551000	12
34	5	0.1601000	11
35	5	0.1651000	11

TABLE 4.1: Number of iterations for γ =5

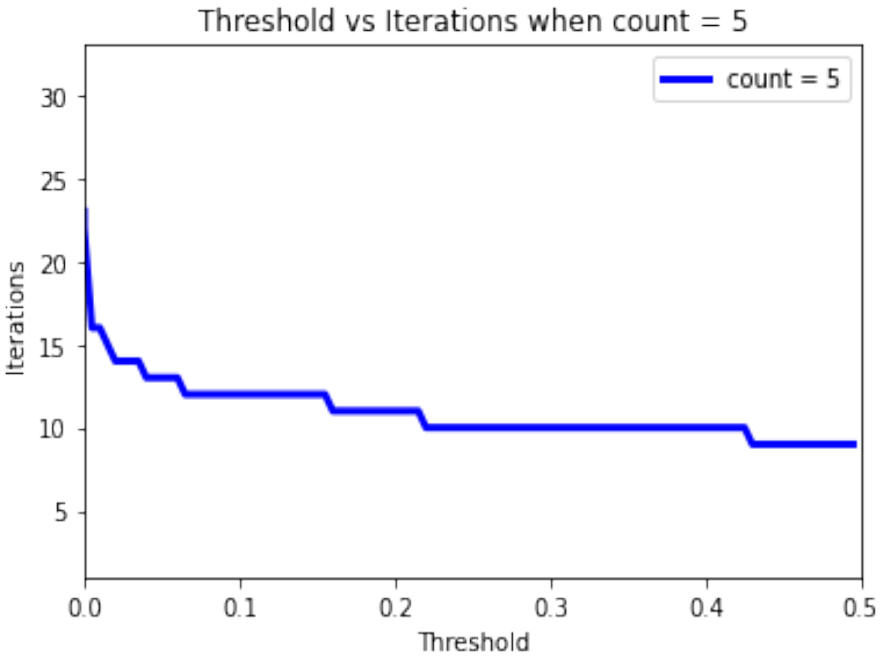


FIGURE 4.1: For $\gamma = 5$

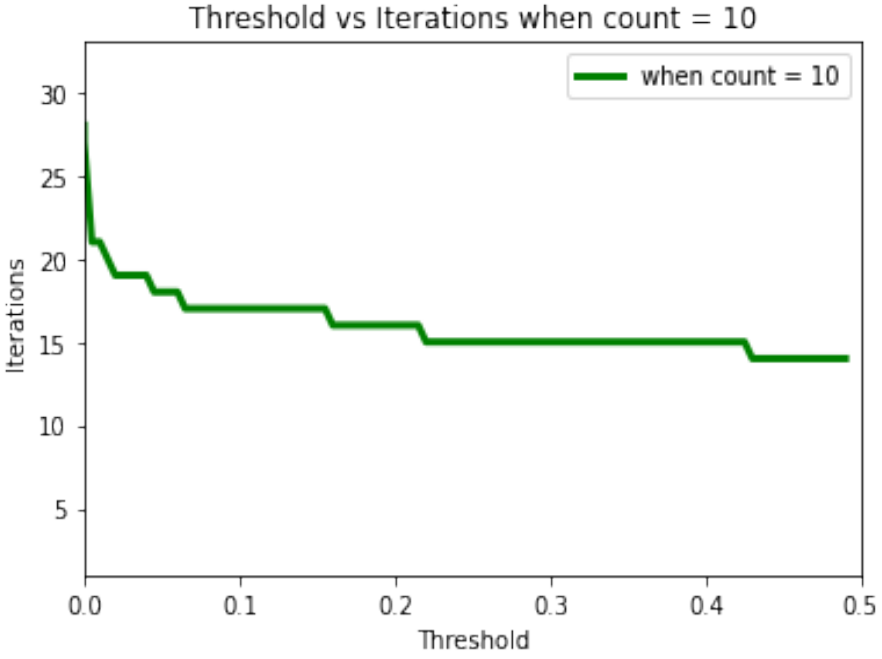


FIGURE 4.2: For $\gamma = 10$

Number of iterations for γ =10			
Serial No	Count of node stability	Threshold	Iteration
1	10	5e-05	28
2	10	0.0001	27
3	10	0.0050	21
4	10	0.0101	21
5	10	0.0151	20
6	10	0.0201	19
7	10	0.0251	19
8	10	0.0301	19
9	10	0.0351	19
10	10	0.040	19
11	10	0.0451	18
12	10	0.0501	18
13	10	0.0550	18
14	10	0.0600	18
15	10	0.0650	17
16	10	0.0701	17
17	10	0.0751	17
18	10	0.0801	17
19	10	0.0851	17
20	10	0.0901	17
21	10	0.0951	17
22	10	0.1001	17
23	10	0.1051	17
24	10	0.1101	17
25	10	0.1151	17

TABLE 4.2: Number of iterations for γ =10

algorithm. The threshold range is from 0.00005 to 0.5 . Figure 4.2 shows that when the threshold is smaller, the number of iterations reduces very slowly. As the threshold increases, the number of iterations does not change much. Even when the threshold is raised, the performance of the algorithm 1 does not suffer. As a result, this method performs well at lower thresholds. We want to use a lower threshold because a larger change in node risk score will not give us an appropriate risk score. That is why we want to select a lower threshold.We see the same findings in the 4.3 for $\gamma = 15$.

The table 4.4 shows that if we increase the threshold, the number of iterations tends to be smaller. If we increase the threshold slowly for fixed γ we can see the number

Number of iterations for γ =15			
Serial No	Count of node stability	Threshold	Iteration
1	15	5e-05	33
2	15	0.0001	32
3	15	0.0050	26
4	15	0.0101	26
5	15	0.0151	25
6	15	0.0201	24
7	15	0.0251	24
8	15	0.0301	24
9	15	0.0351	24
10	15	0.040	24
11	15	0.0451	23
12	15	0.0501	23
13	15	0.0559	23
14	15	0.0609	23
15	15	0.0659	22
16	15	0.0701	22
17	15	0.0751	22
18	15	0.0801	22
19	15	0.0851	22
20	15	0.0901	22
21	15	0.0951	22
22	15	0.1001	22
23	15	0.1051	22
24	15	0.1101	22
25	15	0.1151	22

TABLE 4.3: Number of iterations for γ =15

Iteration for different γ and β					
Iterations No	γ	Lowest Threshold	No of Iterations	Highest Threshold	No of Iterations
1	5	0.00005	23	0.5	9
2	10	0.00005	28	0.5	14
3	15	0.00005	33	0.5	19

TABLE 4.4: For different γ values

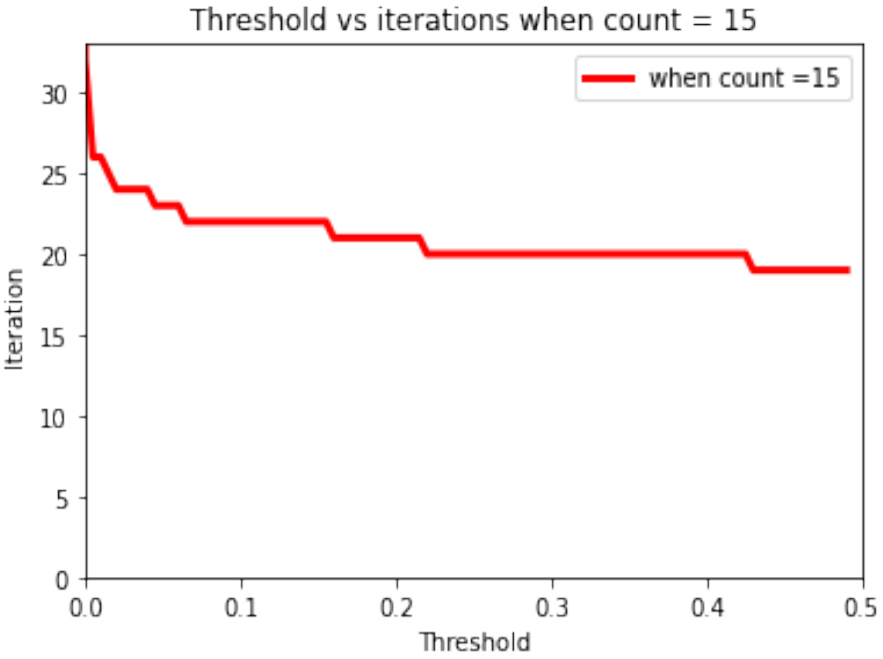
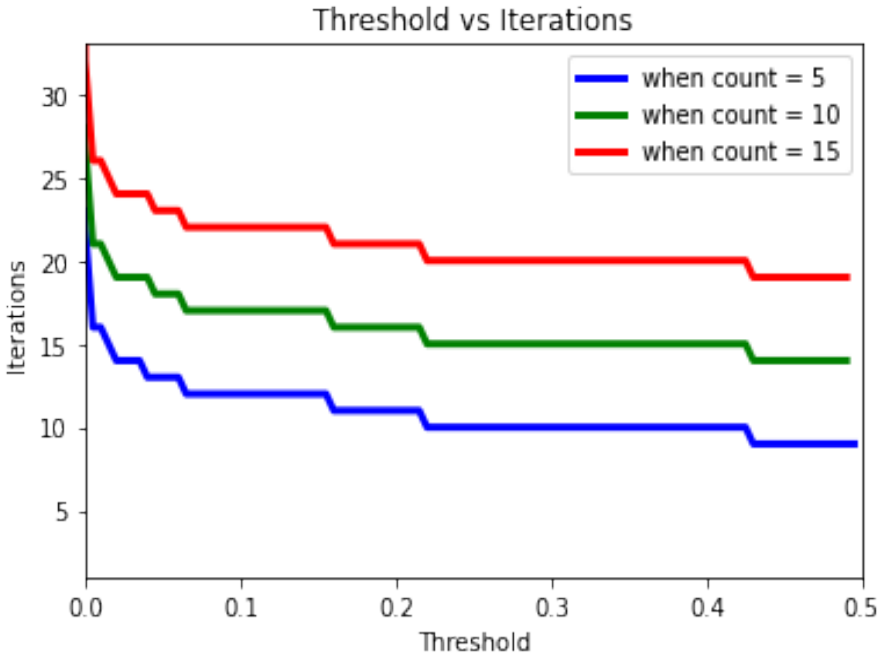


FIGURE 4.3: For $\gamma = 15$



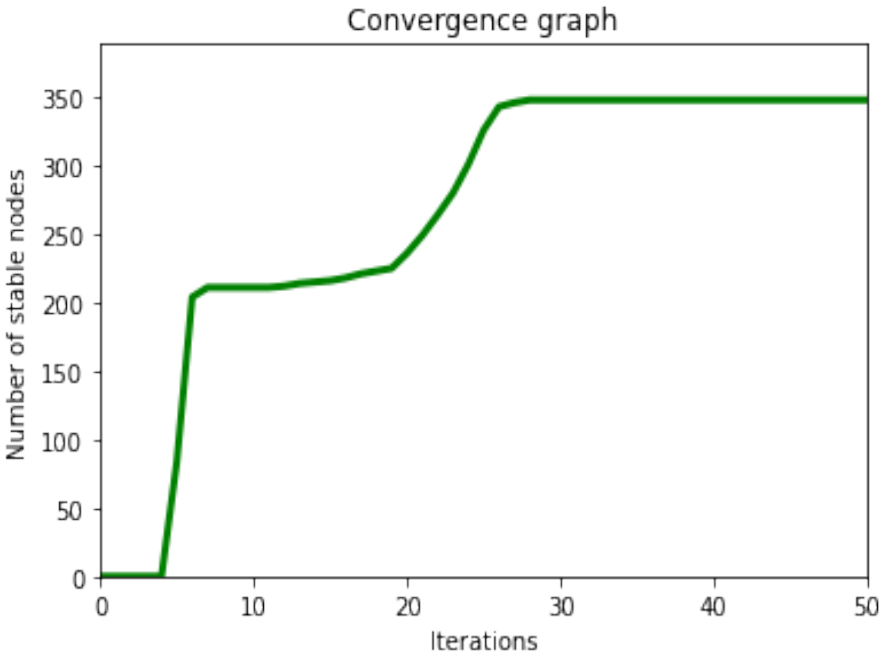


FIGURE 4.4: Convergence graph

of iteration drops at first, but after a certain threshold, the number of iteration decrease very slowly even though we keep increasing the threshold. In polynomial time, the algorithm 1 returns stable risk scores for all parts of the map. As a result, after certain locations become infected, algorithm 1 can efficiently provide us with a result of the risk likelihood of other surrounding nodes.

This graph 4.4 depicts the comparison of stable nodes with each iteration. We can see that algorithm 1 converges very fast. As the range of threshold increases, the execution is completed in a lesser number of iterations.Algorithm 1 returns stable risk score for all parts of the map. As a result, after certain locations become infected, algorithm 1 can efficiently provide us with a result of the risk likelihood of other surrounding nodes.It can help us to get insight on risk probability of each area based on mobility and current infected nodes .

4.4.3 Simulation on the bus dataset

This is the network graph before any risk scores have been assigned. It has 348 nodes and 3328 connections. The initial values of each node are -1. Then we apply

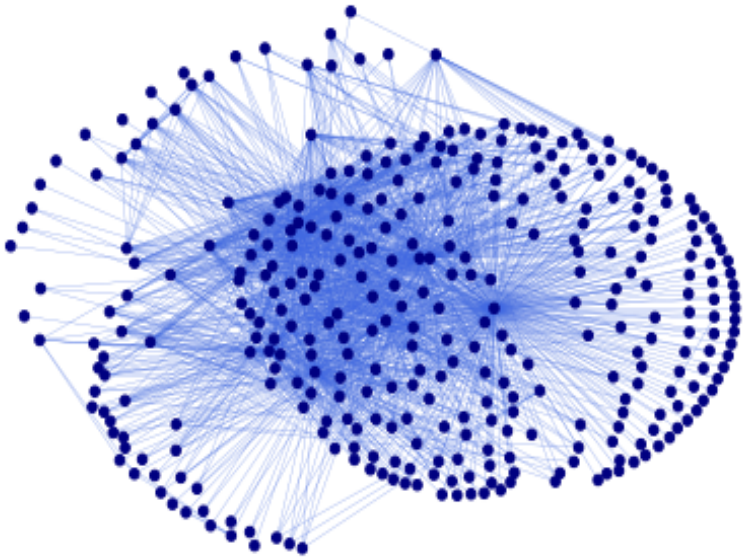


FIGURE 4.5: Initial graph(348 vertices and 3328 edges)

equation 1 to calculate the risk score for the vertices. We assume that four locations (Chowmuhoni , Fakirhat , Patuakhali , Shymnagar) are infected at first, which we depict with the color black. Then the infection propagated to other nodes. Now the mathematical calculations will propagate the score to other neighboring nodes. The formula calculated the risk probability for each of the nodes. Nodes having the highest score are the most vulnerable. Here the lightest red nodes are unlikely to be infected.

4.4.3.1 Threshold and γ values

We select 0.005 as threshold . Threshold indicates the change in risk score of a node before and after an iteration. Every iteration propagates the risk from the neighboring infected nodes to that node. If the threshold is bigger , it can produce unstable results. That's why for this dataset we selected the threshold β to be 0.005 and the value of $\gamma = 5$.

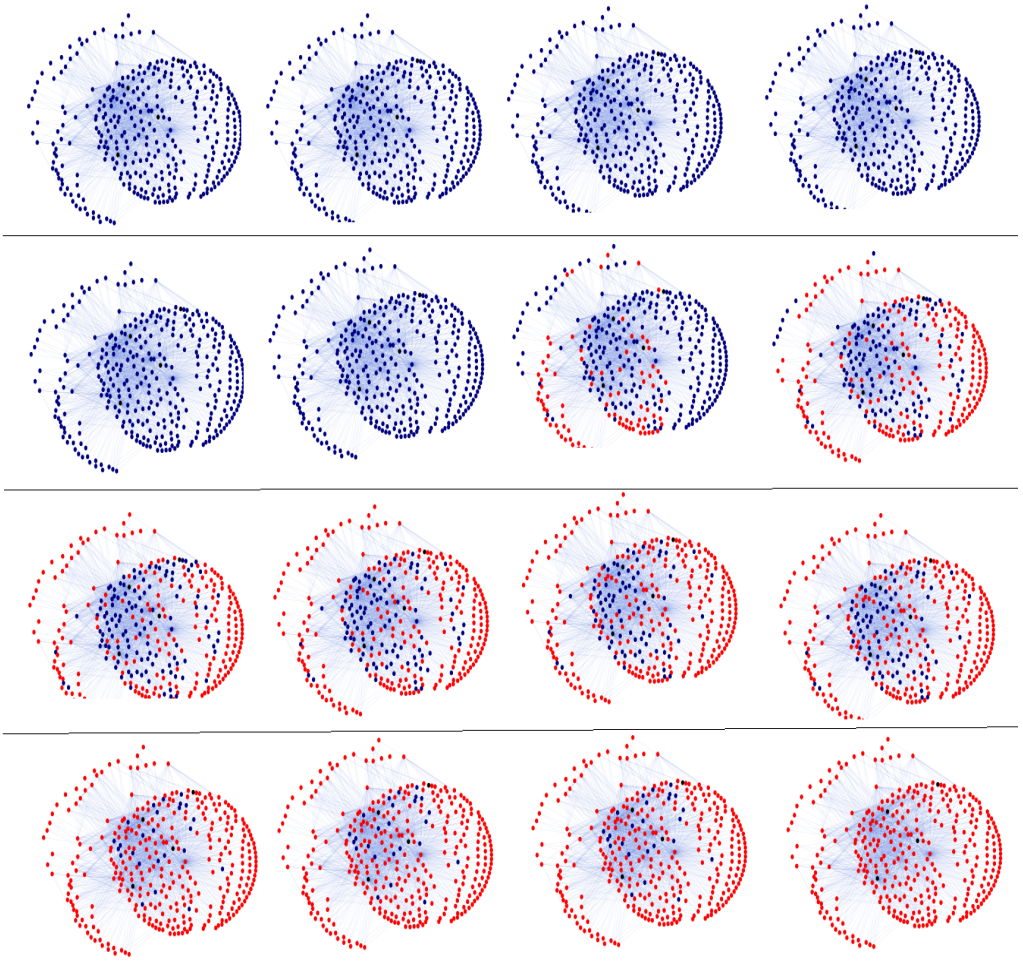


FIGURE 4.6: Iterations

4.4.3.2 Risk score assignment to nodes

Now We set four locations (Chowmuhoni , Fakirhat , Patuakhali , Shymnagar) as infected nodes. These nodes propagate risk to other neighboring nodes with respect to the equation 3.1.

Figure 4.7 shows how the infected nodes(black ones) propagate the score to the other uninfected nodes. With each iteration, the neighboring nodes of infected nodes became infected. After that those infected nodes propagated the risk further. After we get all the risk scores for the vertices, we color-map them.

The figure 4.7 shows that LAMA is the region with the highest risk score of 0.97297.The black nodes in the graph indicate that they were infected first. The

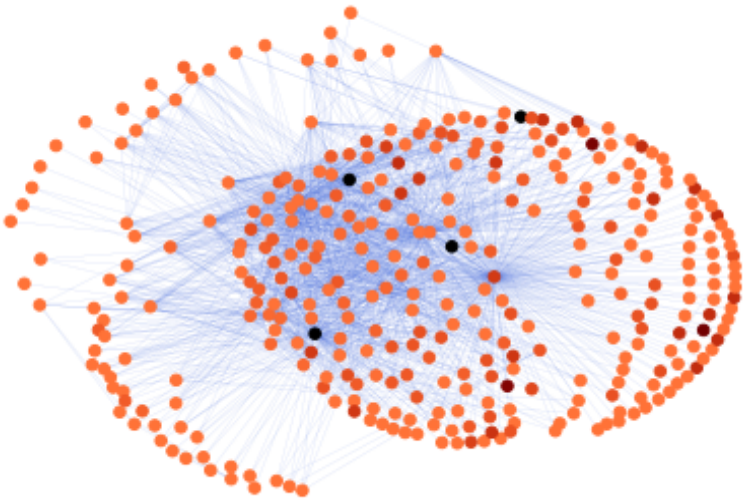


FIGURE 4.7: Scored nodes

light red color nodes are at lesser risk than the dark red nodes in figure 4.7. This helps us to predict the risk probability of different areas of Bangladesh

4.5 Conclusion

In this chapter, we have presented the experimental result of algorithm 1. We calculated the risk probability for different areas considering Bangladesh as a case study. From the result, we can see that algorithm 1 performed very well on the dataset. It gives efficient results on a large dataset with respect to different variable values.

Chapter 5

Conclusions

This chapter presents a discussion of the proposed method and its implication. In addition, it provides a few directions for the researchers as well as our long-term strategy.

5.1 Research Summary

The development of new pathogens is a significant, present-day, and challenging application of methods to predict the risk of diseases. Pathogen emergence and reemergence are happening at alarming rates. In a timely manner in response to significant public health events, the prediction of infectious illnesses can manage widespread outbreaks and decrease the propagation of epidemics. With the use of a system that recognizes the linked zones of a particular site, the infection rate may be greatly controlled.

To control and minimize the damages of the disease we have established a model to calculate the risk score of an uninfected area based on an infected area. As communication has a huge impact on spreading emerging and re-emerging contagious diseases. In our work, we have focused on communication and used it as our parameter. In the proposed system it converts a map of a country into a graph G= (V, E) with the roads as edges, E, and areas as vertices, V. Stable nodes are picked

as the infected vertices. Based on a similarity criterion, other nodes are designated as unstable, and the nodes are progressively checked as stable. As communication plays a huge impact in spreading EID we have set our parameters using mobility data from one area to another. In our experiment, we have used bus communication data [4] and there are 3318 links connecting the 348 places in the converted graph respectively edges and vertices. Before assigning risk scores the value of each node has been set to -1. This risk score algorithm 1 gives values from 0 to 1 where 0 is represented as no risk and 1 an infected node. The higher the risk score of any area the more it is in a risky zone to get infected.

Using the algorithm 1 we have checked the performance where we set different threshold values and get a different number of iterations to get the stability of risk scores of all nodes. From Figure 4.1, Figure 4.2, and Figure 4.3 we can see when the threshold value and iteration number are proportionate. If the threshold value is increased, the number of iterations is also increased to be a stable node.

5.2 Future Work Plan

In this research, we basically work on predicting risk levels for different disinfected areas compared to contaminated areas with the new EID disease using mobility data. So our proposed algorithm 1 is basically based on mobility data. In future, we want to use the combination of factors that have an impact on spreading the disease. We also want to use better communication data

We have already mentioned that our work is based on mobility data and we use bus communication data [4]. We plan to extend this work with vast data. Moreover, we want to extend this work with a Machine learning prediction approach.

Bibliography

[1] Bavinger, J. C., Shantha, J. G., and Yeh, S. Ebola, covid-19 and emerging infectious disease: Lessons learned and future preparedness. *Current opinion in ophthalmology 31*, 5 (2020), 416.

[2] Brachman, P. S. Infectious diseases—past, present, and future. *International Journal of Epidemiology 32*, 5 (10 2003), 684–686.

[3] BROOKES, V. J., HERNÁNDEZ-JOVER, M., BLACK, P. F., and WARD, M. P. Preparedness for emerging infectious diseases: pathways from anticipation to action. *Epidemiology and Infection 143*, 10 (2015), 2043–2058.

[4] BUSbd. Busbd. https://new.busbd.com.bd/.

[5] census. census. http://www.bbs.gov.bd/site/page/47856ad0-7e1c-4aab-bd78-892733bc06eb/Population-and-Housing-Census#.

[6] Changruenngam, S., Bicout, D. J., and Modchang, C. How the individual human mobility spatio-temporally shapes the disease transmission dynamics. *Scientific Reports 10*, 1 (2020), 1–13.

[7] Chiu, H.-Y. R., Hwang, C.-K., Chen, S.-Y., Shih, F.-Y., Han, H.-C., King, C.-C., Gilbert, J. R., Fang, C.-C., and Oyang, Y.-J. Machine learning for emerging infectious disease field responses. *Scientific reports 12*, 1 (2022), 1–13.

[8] Debata, B., Patnaik, P., and Mishra, A. Covid-19 pandemic! it's impact on people, economy, and environment. *Journal of Public Affairs 20*, 4 (2020), e2372.

[9] Findlater, A., and Bogoch, I. I. Human mobility and the global spread of infectious diseases: a focus on air travel. *Trends in parasitology 34*, 9 (2018), 772–783.

[10] Fujiwara, Y., and Irie, G. Efficient label propagation. In *Proceedings of the 31st International Conference on Machine Learning* (Bejing, China, 22–24 Jun 2014), E. P. Xing and T. Jebara, Eds., vol. 32 of *Proceedings of Machine Learning Research*, PMLR, pp. 784–792.

[11] Jocelyne Piret, G. B. Pandemics Throughout History. https://www.frontiersin.org/articles/10.3389/fmicb.2020.631736/full, 2021.

[12] Jung, S., Moon, J., and Hwang, E. Cluster-based analysis of infectious disease occurrences using tensor decomposition: A case study of south korea. *International journal of environmental research and public health 13*, 21 (2020), 5634.

[13] Kanavos, A., Voutos, Y., Grivokostopoulou, F., and Mylonas, P. Evaluating methods for efficient community detection in social networks. *Information 13*, 5 (2022), 209.

[14] Kumar, A., Gupta, P. K., and Srivastava, A. A review of modern technologies for tackling covid-19 pandemic. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14*, 4 (2020), 569–573.

[15] Li, C., Chen, H., Li, T., and Yang, X. A stable community detection approach for complex network based on density peak clustering and label propagation. *Applied Intelligence 52*, 2 (jan 2022), 1188–1208.

[16] Lyzinski, V., Sell, G., and Jansen, A. An evaluation of graph clustering methods for unsupervised term discovery. In *Sixteenth Annual Conference of the International Speech Communication Association* (2015).

[17] Malhotra, D., and Chug, A. A modified label propagation algorithm for community detection in attributed networks. *International Journal of Information Management Data Insights 1*, 2 (2021), 100030.

[18] MayoClinic. Infectious diseases. https://www.mayoclinic.org/diseases-conditions/infectious-diseases/symptoms-causes/syc-20351173, 2022.

[19] Piret, J., and Boivin, G. Pandemics throughout history. *Frontiers Microbio* (2021).

[20] Rees, E., Ng, V., Gachon, P., Mawudeku, A., McKenney, D., Pedlar, J., Yemshanov, D., Parmely, J., and Knox, J. Early detection and prediction of infectious disease outbreaks. *CCDR 45*, 5 (2019).

[21] Sripanidkulchai, K., Rattanaumpawan, P., Ratanasuwan, W., Angkasekwinai, N., Assanasen, S., Werarak, P., Navanukroh, O., Phatharodom, P., and Tocharoenchok, T. A risk prediction model and risk score of sars-cov-2 infection following healthcare-related exposure. *Tropical Medicine and Infectious Disease 7*, 9 (2022), 248.

[22] team, C. Weighted mean. https://corporatefinanceinstitute.com/resources/data-science/weighted-mean/.

[23] Wang, F., and Zhang, C. Label propagation through linear neighborhoods. In *Proceedings of the 23rd international conference on Machine learning* (2006), pp. 985–992.

[24] Win, H. N., and Lynn, K. T. Community detection in facebook with outlier recognition. In *2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)* (2017), IEEE, pp. 155–159.

[25] Woolhouse, M. How to make predictions about future infectious disease risks. *Philosophical Transactions of the Royal Society B: Biological Sciences 366*, 1573 (2011), 2045–2054.

[26] Yang, W., Zhang, J., and Ma, R. The prediction of infectious diseases: A bibliometric analysis. *International Journal of Environmental Research and Public Health 17*, 17 (2020), 6218.