UNIVERSITY OF OSLO
Department of informatics

Epidemic Network and Centrality

Master thesis

Akram H. Rustam

May 2006

Abstract

This project is about epidemics spreading in computer networks and the issue of node centrality. The aim of such analysis is to investigate the rate of infection and information spreading in the network, to find the most important nodes in the network graph, and finally to answer the research question which states that *"centrality of the node has a crucial role on spreading power"*.
The method used in this project to answer the research question is important because it measures the power of spreading information by one specific node and studies the environments around it, instead of environments around the whole network. So by finding the power of spreading and properties of one specific node in the network will help us understand which weaknesses or advantages this node has for maintenance or blocking hazards at the right time.

The position or location of each node in the network is studied in a form of degree, betweenness, and centrality of the node and the rate of effect those properties have on spreading of information.

Hypotheses are suggested on epidemic networks in addition to our research question and graphs are generated and analyzed, to test those hypotheses. We do so by developing a mathematical SI-model which is depending on the values of principal eigenvector to measure the number of infected nodes as a function of time, and also trying to monitor infections' movements and expressing the frequency and cumulative tables, and graphs to support and confirm our developed mathematical method.

The obtained results from our work in this thesis show that centrality of the node is related to the power of information spreading.

Acknowledgements

First of all I would like to thank my supervisor Kirsten Ribu for her valuable and useful guidance to fulfill this project. Second, I would like to thank Professor Mark Burgess for his special effort through the Masters course, and I would also like to thank Kyrre Begnum, Hårek Haugerud, Simen Hagen, Tore Møller Jonassen, and Siri Fagernes for their support through the Masters course.

Furthermore, I would like to thank my classmates specially Ilir Bytci and my friend Raheel A. Chaudhry for their ever support during the two years period of master program.

Finally I would thank my wife for her support and patience during my study in despite of she was busy with her college studies and we have a daughter to take care of.

Preface

This Master thesis is written to be as a partial fulfillment of the requirements for the degree of Master of Science in Network and System Administration
at Oslo University College,
2006.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 2
- 1.2 Aims of this thesis . . . 3
- 1.3 Expectations and hypothesis . . . 3
- 1.4 Contribution . . . 4
- 1.5 Limitations . . . 4

2 Background **5**
- 2.1 Network graph . . . 5
- 2.2 Network topologies . . . 8
- 2.3 Centrality . . . 9
- 2.4 Principal Eigenvector . . . 9

3 Related work **13**
- 3.1 Previous work on epidemic network . . . 14
  - 3.1.1 Epidemic Models . . . 14
- 3.2 Epidemic networks and centrality . . . 23

4 Methodology **31**
- 4.1 Research Subjects . . . 31
- 4.2 Research Tools . . . 31
- 4.3 Procedures . . . 32
- 4.4 Proportion . . . 33
- 4.5 Mathematical method . . . 36
- 4.6 Graphic Representations of Data . . . 41
  - 4.6.1 Tracing infections' movements . . . 42
  - 4.6.2 Cumulative Representation . . . 42

5 Results and Analysis **44**
- 5.1 Result from our small network . . . 44
  - 5.1.1 Mathematical Method applied on small network . . . 46
  - 5.1.2 Tracing infections' movements for small network . . . 49
- 5.2 Result from our large network . . . 62
  - 5.2.1 Mathematical Method applied on large network . . . 63

5.2.2 Tracing infections' movements for large network 64

6 Conclusions and Discussion **70**
6.1 Conclusions 70
6.2 Results Utilities and Recommendations 72

Appendices **73**

A Adjacency matrix **74**

B Report output **76**

C Bar chart graphs **79**

D Tracing infections' movements **83**

E Eigenvector Ranking **86**

List of Figures

2.1 The structure of Internet at (a) the router level and (b) the inter-domain level 6
2.2 a) Directed b) Undirected 7
2.3 Network topologies 8
2.4 More network topologies 8
2.5 The most central node is B. 9
2.6 A very simple network 10

3.1 Random Scanning of Active Worms 14
3.2 SI model 16
3.3 SIS model 19
3.4 SIR model 21
3.5 SIDR model 22
3.6 SIRS model 23
3.7 Nodes A and B have highest betweenness than node C 24
3.8 Curve of rate of infection (S-Shape) 27
3.9 Cumulative S-shape with $p = 0.05$ 28
3.10 Cumulative S-shape with $p = 0.6$ 28

4.1 Network consist of 10 nodes 34

5.1 Small network consists of 12 nodes 44
5.2 screenshot from application of our developed SI-model for all $nodes$ in graph of Figure 5.1 47
5.3 Screenshot from application of our developed SI-model for $node_2$ and $node_6$ 48
5.4 Tracing infections' movements from $node2$ 50
5.5 The Frequency and Cumulative graphical representation for Table (5.2) 52
5.6 Tracing infections' movements from $node6$ 53
5.7 The Frequency and Cumulative graphical representation for Table (5.3) 54
5.8 The cumulative graphical representation for Table (5.4) 55
5.9 Histogram quantities representation of nodes betweenness, eigenvector and degree for Table (5.1) 56
5.10 Bar chart quantities representation of nodes betweenness 57
5.11 Bar chart quantities representation of nodes degree 58
5.12 Bar chart quantities representation of nodes eigenvector 59
5.13 Scatter quantities representation of nodes eigenvector versus betweenness. 60
5.14 Scatter quantities representation of nodes eigenvector versus degree. . 61
5.15 Network of 100 nodes 62
5.16 Screenshot of mathematical application of our SI model for all nodes in graph of Figure 5.15 63
5.17 Screenshot of mathematical application of our SI model for $node44$ . . 64
5.18 The Frequency and Cumulative graphical representation for Table (5.5) 65
5.19 The Frequency and Cumulative graphical representation for Table (5.6) 66
5.20 The cumulative graphical representation for Table (5.7) 68

C.1 Bar chart quantities representation of nodes betweenness from large graph 80
C.2 Bar chart quantities representation of nodes eigenvector from large graph 81
C.3 Bar chart quantities representation of nodes degree from large graph . 82

D.1 Tracing infections' movements from $node44$ 84
D.2 Tracing infections' movements from $node57$ 85

List of Tables

5.1 Betweenness, Eigenvector, and Degree output from ORA Risk Report 45
5.2 Frequency and Cumulative table for infecting other nodes by $node2$ at t_i 51

5.3 Frequency and Cumulative table for infecting other nodes by $node6$ at t_i 53
5.4 Cumulative table for infecting other nodes by $node2$ and $node6$ for our small network (12 nodes) 54
5.5 Frequency and Cumulative table for infecting other nodes by $node44$ at t_i 65
5.6 Frequency and Cumulative table for infecting other nodes by $node57$ at t_i 66
5.7 Cumulative table for infecting other nodes by $node44$ and $node57$ in our large network (100 nodes). 67

B.1 Betweenness, Eigenvector, and Degree output from ORA risk report . 78

Chapter 1

Introduction

Recently when Internet has grown and become available for every one and e-business also has become incrasingly popular; a natural phenomenon to appear was the concept of the so-called epidemic network. The term of epidemiology is used indeed for human diseases for a long time now. Epidemiology [1]is dealing with disease spreading within populations and can be defined as *"the science of the infective diseases - their prime causes, propagation and prevention. More especially it deals with their epidemic manifestation" (LeRiche & Milner, 1971).*

And since some computer worms propagated themselves in a very high speed and rapidly such as Code-Red and SQL Slammer and that propagation can be described by epidemic models as those that have been used for biological epidemiology [2]. Where hosts (computers) can be considered as a population and some of them are infected and contagious which can infect other susceptible hosts in the population by some infection parameter β.

So we can consider a network of machines (router, set of routers, hosts) as any other population which can get epidemiological diseases from each other, unless that the main different is within technological world, the spreading of information (worms, viruses, etc.) will not take effect without contact between individuals.

Viral attack ca be contained by using antivirus programs or human countermeasures but often when a new epidemiological worm spreads may it will be difficult to detect it and contain it at once, and this will causes too much damage and thousands of machines (hosts) will be infected. On july 19th, 2001 a worm (*"Code-Red v2"*) was spread into the internet and infected around 360000 machines over 14 hours and that cost almost $2.6 billion [3].

Thus it is necessary to study the topology of the network and if an important node is infected how this shall infect other nodes due to its central role. So having a good knowledge about the structure of the network is very necessary for maintenance and security purpose.

Since mathematical methods can give us a clear view and can be a good help to identify and solve many complex problems that why we would like to develop a new mathematical method for information spreading by using principal eigenvector values (PEV) and then analyze and determine that how eigenvector and betweenness centrality of the nodes within the network is related to the rate of infection.

The outline of this thesis will be like this: first we shall review background in form of network topologies, network graphs, and principal eigenvector and centrality. Then we shall talk about previous work has been done by others and review some known epidemic models for spreading of infection.(We used the term ”spread of infection” and ”spread of information” interchangeably).Finally we shall explain our methodology to fulfill our study and then followed by result analyzing and conclusion.

1.1 Motivation

After joining the course of (Analytical Network And System Administration) coordinated by Professor Mark Burgess at Oslo University College, autumn 2005, I found interest in the network structure and especially the ways to rank all nodes in the network.

Ranking nodes done by calculating adjacency matrix (see Section 2.1 and 2.4) and then finding principal eigenvector which it’s values represent nodes centrality in the network and highest value indicates the most important node. This knowledge led me to be increasingly curious to study more about principal eigenvector. Our motivation was trying to gain more advantage from principal eigenvector and determining how eigenvector and betweenness centrality of the nodes is related to the rate of infection in epidemic network.

Later when we had to choose our final thesis for obtaining the degree of Master of Science in Network and System Administration and after discussion with Professor Mark Burgess I saw that it was a good opportunity to choose a thesis which deals with epidemic network since it includes implicit working with network structure and relationship between nodes in form of nodes degree and position (centrality) and their effect on infections behavior.

1.2 Aims of this thesis

The aims of this thesis are as follows:

1. Formulate a hypothesis which investigates centrality and rate of infection and then trying to make graphs that test whether it is true or not.
2. Getting more knowledge and understanding about epidemic network and centrality.
3. Figure out important nodes.
4. Observing the behavior of the information spreading by one chosen node.
5. Answering research questions.

1.3 Expectations and hypothesis

Our Research Question: Does principal eigenvector and centrality of the nodes related to the rate of information (infection) spreading?

Our hypothesis:

Principal eigenvector and centrality of the nodes is related to the rate of infection (spreading of infection). So if we suppose that there is no recovery during the infections period; a node with highest principal eigenvector value (PEV) which is called *"most important node"* shall infect all other nodes rapidly than any other nodes and this lead to rapid growth in curve of infection. Thus principal eigenvector is a good measurement for centrality of the nodes.

Furthermore we have a prediction which represents what we expect from our developed mathematical method for epidemic information spreading:

- Nodes with different centrality have different curves of infections regardless of their degree average.

1.4 Contribution

To answer research question and to show that whether our hypothesis is true this thesis provides two main contributions:

- First is a developed mathematical SI-model depending on (PEV) for each node. This method consists of three main phases:
  1. Generating a random network of nodes by usin ORA see section (4.2) which nodes interacting with each other.
  2. Finding principal eigenvector (PEV) for each node to be assumed as rate of infection (τ).
  3. And finally executing phase by applying our developed mathematical SI-model which is represents the curve of infection as a function of time see Eq.(4.32).
- Second is tracing infection' movements to support and confirm our mathematical method. This method also consists of three main phases:
  1. Finding the number of infected nodes at each unit time (Frequency).
  2. Cumulative phase by summing all infected nodes at each unit time.
  3. Making statistical graphs depending on Frequency and Cumulative values to show the curve of infection.

1.5 Limitations

Our method is not for describing one special worm (virus) spreading such as RedCod, or SQL Slammer but our method shall describe information spreading in general by one infected node depending on centrality and principal eigenvector (PEV).

Our mathematical SI-model:

$$\boldsymbol{I(t) = \frac{(N-1)}{1+(N-2)e^{-\tau(N-1)t}}}$$

this equation represents the curve of infection see section (4.5) and can be used for $N = n$ nodes where $n = 2, 3, ...\infty$.

Chapter 2

Background

2.1 Network graph

Network graph consist of nodes which are connected to each other by edges, i.e links. Before we start to talk about properties of nodes and links we should define each of them to be easier to understand later in the following chapters.

Network can be of two types [4]: physical and logical (virtual). Physical network where for example computers with assigned different IP addresses interacting with each other according some protocols, and this kind of network can be represented by undirected graph, where nodes in the graph represents the computers and the edges between nodes as the physical links (wire, optical cable, etc) for communications. So the connection between node i and node j happens through the path between node i and node j. Another type of network graph is logical or virtual network such as e-mail graph, where each node in this graph represents a user and each link from node i would go to all other users (nodes) which have e-mail address in the e-mail address book of node i, i.e. link represents contact between users (nodes). And similar to that in the web graph node represents a web page and each link represents a hyperlink. All these types of graph [4] (undirected graph, e-mail graph, and web graph) have almost the same properties when it considers degree of nodes where [5] *"the graph of Internet is sparse with 75% of the nodes having outdegrees less or equal to two"*, grouping or clustering coefficient, and distance average *"distances between any connected nodes"* and the distance between two nodes [5] is the sum of all links of the shortest path between them.

Furthermore the Internet [5] can be divided into *"subnetworks"* which are called *"domains or autonomous systems"* and these *"subnetworks"* interacting or connecting to each other by different administrative authorities. And according to [5] there are two levels Internet graphs namely *"router level"*see Figure (2.1) (a) where each router (black dot) represented by a node and *"inter-domain level"* where each *"domains or autonomous systems"* see Figure (2.1) (b) represented by a single node and link between routers inside one domain is represents inter-connection.

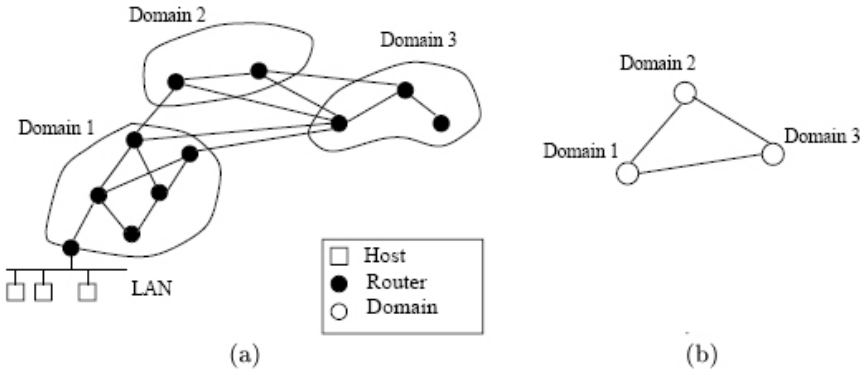


Figure 2.1: The structure of Internet at (a) the router level and (b) the inter-domain level

And also according to [4] there are two levels of Internet graph: *"microscopic Internet graph"* in this kind of graph the node represents routers and hosts and link represents communication between them. And second is *"macroscopic Internet graph"* in this graph the node represents an *"Autonomous System"* which consist of a set of routers and link represents communication between *"Autonomous System"*. Two nodes *"Autonomous System"* in macroscopic graph are adjacent if there are at least two routers which can communicate with each other between those two nodes.

Thus node can represents computer, user, webpage, host, router, and subnetwork (domain) and links can represents physical material such as wire and optical cable for communication between nodes or can represents hyperlink to connect webpages or represents contact between users or routers or subnetworks.

Nodes can be represented [6] as dots they are connecting to each other by lines with or with out arrow on them, and these diagrams called *"graph"* as D. Kønig proposed it.

Each line can indicate to some property such as:

- *"A dominates B (directed)"*
- *"A depends on B (directed)"*
- *"A is associate with B (undirected)"*

Where directed link means *"one-way"* and undirected means *"multi-way"*. Each node in the graph has its degree that depends on the nearest neighbors. And degree of a node can be defined as: *"In a non-directed graph, the number of links connecting node i to all other nodes is called the degree k_i of the node. In a direct graph, we*

distinguish incoming and outgoing degrees"[6]. If we look at Figure:2.2 b) that each node has degree of 2 (k=2).

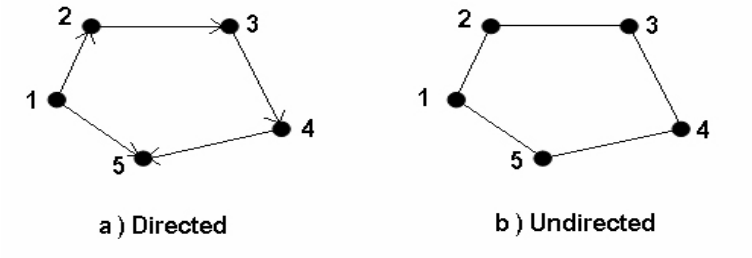


Figure 2.2: a) Directed b) Undirected

We can represent any graph by *"adjacency matrix"* to be easier to deal with and managed. Adjacency matrix is contains of 0's and 1's: where "1" indicates connection between the concerned nodes and "0" indicates no connection. In adjacency matrix the number of rows is equal to the number of columns and labeled by nodes of the graph.

Here we can represent Figure (2.2) b) by an adjacency matrix as in Eq. (2.1):

$$A = \begin{array}{c} \\ node1 \\ node2 \\ node3 \\ node4 \\ node5 \end{array} \begin{array}{c} \begin{array}{ccccc} node1 & node2 & node3 & node4 & node5 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \end{array} \tag{2.1}$$

2.2 Network topologies

Networks can be represented by many types of graphs, that according to their connectivity among nodes. As [6] shows three most important types of network topologies which is discussed by Paul Baran in 1964, namely as in Figure:2.3 *"(a) centralized, (b) de-centralized, and (c) distributed mesh"*.

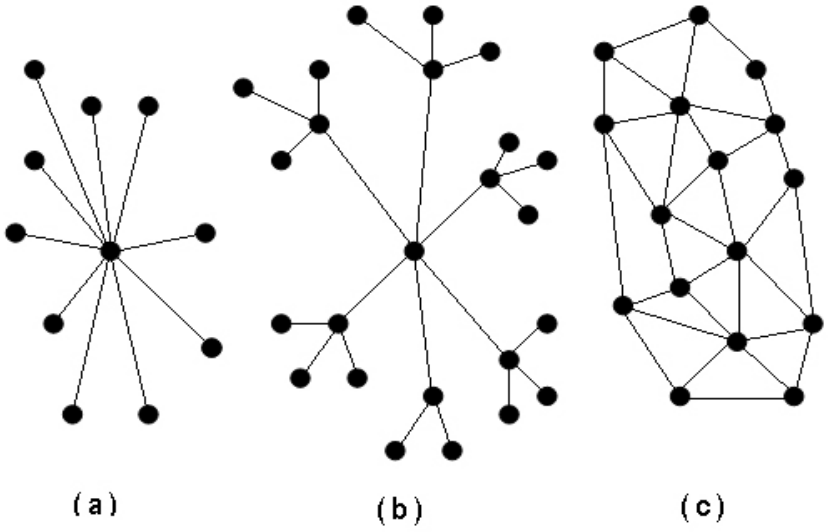


Figure 2.3: Network topologies

There are other topologies such as in Figure:2.4) (a) bus (line), (b) ring, (c) wheel, and (d) grid.

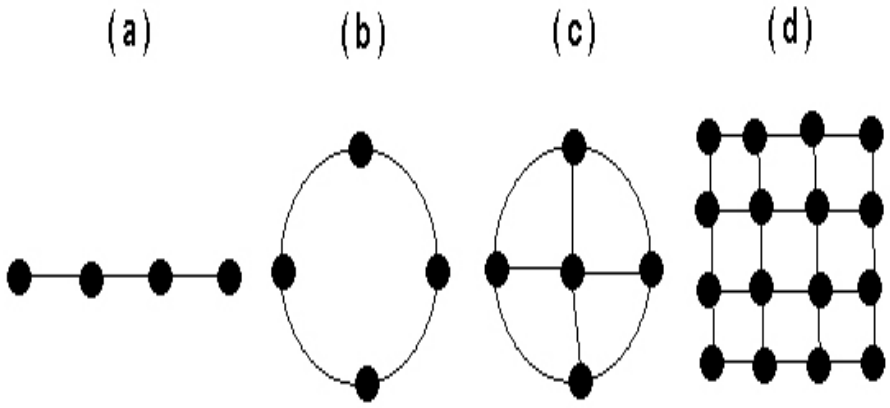


Figure 2.4: More network topologies

2.3 Centrality

Centrality is one of the most important properties of the network analysis. If a node has central position will has a crucial role to spread information, or will be one of the most dangerous point which we should deal with by care. Those nodes are *"well-connected"* with others and they have contact with many other important nodes [6].

Not all nodes have the same level of effect even may they have the same degree that because of their position within network [6]. As we can see from the Figure (2.5) both node A and node B have the same degree but node B is more important than node A because of its position in the network which lies between many important nodes.

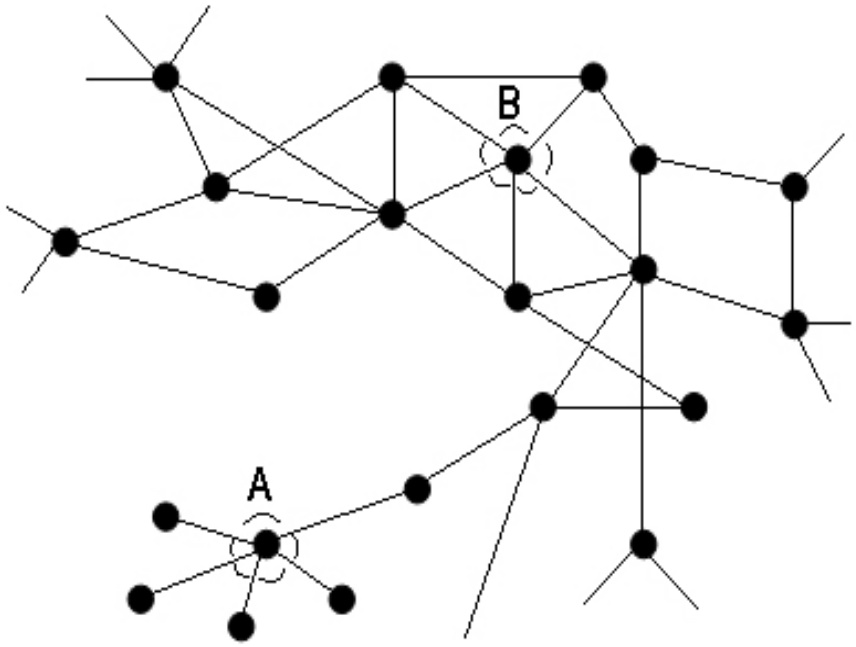


Figure 2.5: The most central node is B.

2.4 Principal Eigenvector

Since we want to depend on power of principal eigenvector values (PEV) in our method to relate these values as probability of the rate of information spreading, so we would like to review in general what is principal eigenvector and what principal eigenvector is used for especially within network.

Suppose A is an $N \times N$ adjacency matrix then this will produce N eigenvalues and N eigenvectors. We will choose the highest eigenvalue of that matrix to calculate the

principal eigenvector, and most central node has highest value which is represent the *"eigencentre"* of the graph, and all values in this eigenvector are positive.

Suppose that we have a very simple network which consists of just three nodes as shown in Figure (2.6):

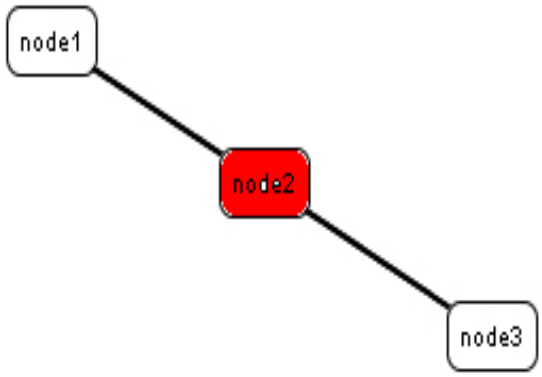


Figure 2.6: A very simple network

Figure (2.6) can be represented [6, 7] by an adjacency matrix as in Eq. (2.2):

$$A = \begin{array}{c} \\ node1 \\ node2 \\ node3 \end{array} \begin{array}{c} \begin{bmatrix} node1 & node2 & node3 \end{bmatrix} \\ \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \end{array} \tag{2.2}$$

Where A is an $N \times N$ adjacency matrix and we can find eigenvalues and eigenvectors by applying Eq. (2.3):

$$A\vec{v} = \lambda\vec{v} \tag{2.3}$$

Where λ is called eigenvalues and there are correspondingly solutions which is called eigenvectors. Each eigenvector can be represented as $N \times 1$ matrix as Eq.(2.4):

$$\vec{v} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{2.4}$$

Eq.(2.3) can be written as:

$$(A - \lambda I)\,\vec{v} = 0 \tag{2.5}$$

Where I is an identity matrix and has the same dimensions as A.

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

By setting

$$\left|A - \lambda I\right| = 0 \tag{2.6}$$

Then Eq. (2.6) gives

$$\begin{vmatrix} -\lambda & 1 & 0 \\ 1 & -\lambda & 1 \\ 0 & 1 & -\lambda \end{vmatrix} = 0 \tag{2.7}$$

And by some algebra calculation we will get eigenvalues(λ) = (0, $\sqrt{2}$, -$\sqrt{2}$).

Now we are not interested to find all eigenvectors wich are corresponding to eigenvalues but only the principal eigenvector, so we will choose the highest value of λ ($\sqrt{2}$)

then we rewritethe Eq. (2.5) as:

$$\begin{bmatrix} -\sqrt{2} & 1 & 0 \\ 1 & -\sqrt{2} & 1 \\ 0 & 1 & -\sqrt{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0 \tag{2.8}$$

And this gives the principal eigenvector:

$$\vec{v} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ \sqrt{2} \\ 1 \end{bmatrix} \tag{2.9}$$

Eq.(2.9) is an eigenvector corresponding to the eigenvalue $\sqrt{2}$.

The equation $\vec{v} = [1, \sqrt{2}, 1]^T$ represents principal eigenvector for our network (Figure 2.6)and these values characterizing the figure in some way that node2 with highest eigenvector value $\sqrt{2}$ is most important node in our network. We can see that from the Figure(2.6) if node 2 goes down the whole network will goes down, but if node1 or node3 goes down still we have connection between node2 and node3 or node2 and node1.

Chapter 3

Related work

First this chapter reviews some most known epidemiological models such as SI, SIS, SIR, SIDR, and SIRS models and then reviews related works have been done on epidemic networks and centrality.

Since network [8, 9] consists of nodes (vertices or hosts) which can be represented by graph, we can consider the contact between the nodes as the edge which let the information (diseases) pass through or transmitted. Nodes are connected to each other, and the number of contacts represents the degree of the node. Not all the nodes have very high degree unless complete graph which all nodes have equal degree and those nodes are considered as most important nodes. Since network structure has a biggest role of information spreading, therefore we should give a careful attention to the degree distributions among nodes because of their role of building the network structure.

It is useful to know something about worm spreading because most epidemic models talking about infection by worms or viruses. When a worm [10] spreads in the Internet trying to infect the vulnerable machines and after the infection happens those vulnerable machines will get a copy of that worm. And by the same way these new infected machines tries to infect other machines and so on.

The attacker try to make a list of machines with high network connection which is called *"hitlist"*, first the worm will begins to infect those collected machines down the list, and then these machines will infects other vulnerable machines.

The worm spreading mechanisms are many for instance *"random, local subnets, permutation, and topological scanning"*.

Each computer will try to infect others in the Internet according to random scanning, as in fully connected network and each node in this kind of graph represents a computer and each link represents a connection. In subnet scanning there are direct contacts between computers in this case the worm will not scanning randomly, but instead scans for hosts on the local address space. And according to [10] if some machine gets infected this machine will not be infected again.

And by applying random scanning of active worms one can see the rate of infection as in Figure (3.1) which is showing the number of infected nodes versus the time.

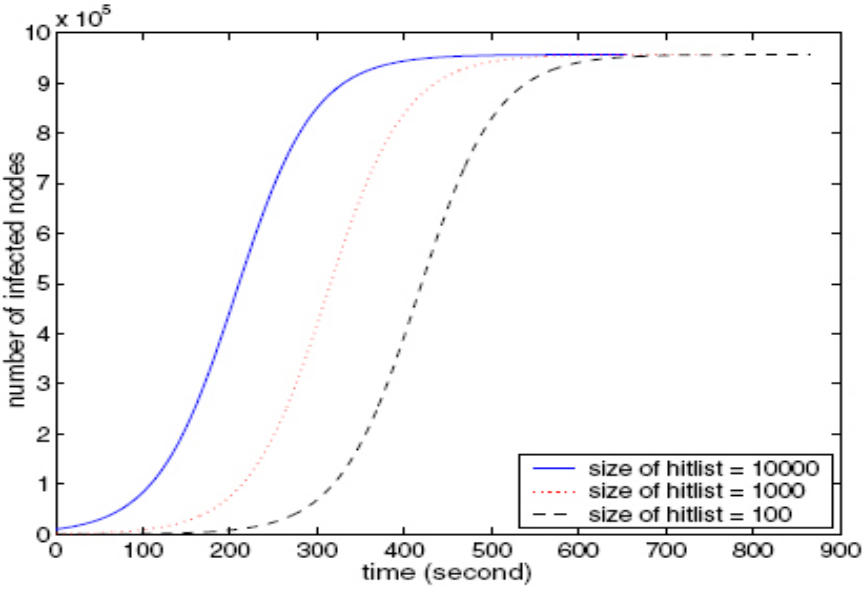


Figure 3.1: Random Scanning of Active Worms

3.1 Previous work on epidemic network

3.1.1 Epidemic Models

Epidemic model [11] is a good tool to understand the information (disease) spreading by relating the process of spreading to the individuals (hosts) properties.
Anyway epidemic models are not easy to apply and to be sure of their results because of:

1. Their conclusions depend on assumptions which are rarely straightforward.
2. Some times they can fit the date virtually to their models easily because the epidemic threshold is very strong which is easy to be observed.
3. Depending on parameter values; such as number of population and units (rates) of the contact between them which are just assumptions.

To make the epidemic process possible or easier it should not be complex and should be very simple and clear to understand.

Epidemic algorithm deals with population that can be represented by a set of individuals which interactive with each other according some rules and these rules have a crucial role to spread the information.

Those individuals should have one of these states at a specific time [3, 12]:

1. Susceptible:
   The individual has no idea about the specific information (worm, virus, etc.), but has ability to get that information.

2. Infective:
   The individual is knows about the specific information and will infect others by spreading that information that means they are vulnerable but not become victims yet.

3. Recovered:
   The individual knows about that specific information but will not infect any other.

There are to useful models for infectious disease, the first is stochastic and the second is deterministic [13, 9].

Stochastic models uses for small or isolated population depending on chance by following each individual and the number of infected nodes converges to zero this means the extinction of worm happens with probability one. Stochastic models need much work to get a result which confirms the predictions. Also these models can be difficult to understand and complex mathematically.

Deterministic models uses for large population, trying to tell us what happened to the average of population insteady state by deciding some initial condition. These models put the individuals in subclasses or states. For example SEIR model includes these states: Susceptible, Exposed, Infected, and Recovered. Deterministic models uses widely because they do not need too much data and they are not complex.

Transition from one state to other happens at some rate, for instance infection rate is very well known factor which force susceptible individuals to change their state to infectious individuals.

When an epidemic disease appears and because the individuals (population) interacting with each other in a particular situation, the individuals will change their states by time. And transition from one state to other happens by some rate for instance infection rate is very well known factor which force susceptible individuals to change state to infectious individuals.
At start each individual can be considered as susceptible (S), then by time the number of susceptible individuals will decrease and the number of others (infected (I), exposed (E), and recovered (R)) will increase by some chosen rates.

Recently epidemiological models for network become more and more popular for virus and worm propagation. Network can be represented by graphs [4, 14, 15], and each graph consists of nodes which represent individuals and edges (links) which represent the possible contacts between the individuals. Each node in the graphs has one

of these states: infectious, susceptible, exposed, recovered, and removed. Any infected node can pass the infection to its susceptible neighbor node.

There are several factors that cause or influence the spread of an infection [3, 16, 17]:

1. The number of infected nodes at the present time.
2. The rate of infection.
3. The number of susceptible nodes.
4. The rate of infection or transmission.
5. The vulnerability of the population.
6. The immunity levels.
7. The state which any worm can be ready or prepared for copying it self.
8. The period of time that one infected node can stay infected.
9. Degree of connection with other nodes.

SI model

SI model considers as one of the simplest epidemic model to describe the growth of an infection. The individuals (nodes) divided in to compartments or states: Susceptible (S) and Infectious (I) [3, 12].

In this model nearly each individual is susceptible. After spreading of information (disease) all individuals (susceptible) will be infected exponentially and will remain infected.

This model assumes [12, 4] that: first the infected individuals will remain infected for ever, that means there is no birth, latency, death, or recover among them, second the population size is large and fixed, and third the population is homogeneous.

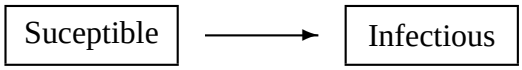


Figure 3.2: SI model

This model [4] can be used for *"worst-case propagation"*, i.e. there are not any security protections is used such as automated network security (antivirus, firewalls, intrusion detection system, monitoring, etc.) and no countermeasures action taken to protect against worm propagation such as (traffic blocking, patching, etc.).

"While computer worms represent a relatively new threat, the mathematical foundations governing the spread of infectious disease are well understood and are easily adapted to this task."[3].

So SI model can be described mathematically by the differential equation [3, 4].

$$\frac{d_i(t)}{\mathrm{d}t} = \beta\bar{d}(1 - i(t))i(t) \tag{3.1}$$

Total rate of newly infected nodes

Where:

β	is a rate of infection and it is an assumed constant.
$\bar{d}$	is the average degree of an infectious node.
$i(t)$	is fraction of infectious node (I(t)/N) at time t.
N	is population number
$I(t)$	infectives nodes (spreading the infection) at time t.
$\bar{d}(1 - i(t))$	is the expected number of susceptible neighbors which can be infected by an infectious node.
$\beta\bar{d}(1 - i(t))i(t)$	is the total rate of infected nodes.

The solution to Eq.(3.1) is:

$$i(t) = \frac{i(0)e^{\beta' t}}{1 - i(0) + i(0)e^{\beta' t}} \tag{3.2}$$

Logistic curve describing the rate of infection

Where: $\beta' = \beta\bar{d}$

This equation will give the S-shape curve [4], which indicates that there are three regions of fraction of infection nodes:

1. *"Slow start:"* in this phase there are not many nodes infected yet.

2. *"Exponential growth:"* in this phase the number of infected nodes will increase exponentially.
3. *"Equilibrium:"* in this phase the number of infectious nodes will change gradually.

We can rewrite Eq (3.2) as:

$$i(t) = \frac{i(0)e^{\beta \bar{d} t}}{1 - i(0) + i(0)e^{\beta \bar{d} t}} \tag{3.3}$$

Logistic curve describing the rate of infection

Thus the equation for spreading infection on complete graph where $\bar{d}$ (average of degree) = $(n-1)$ will be as Eq (3.4):

$$i(t) = \frac{i(0)e^{\beta (n-1) t}}{1 - i(0) + i(0)e^{\beta (n-1) t}} \tag{3.4}$$

Logistic curve describing the rate of infection for complete graph

And SI-model [18]can be represented mathematically as:

$$\frac{da}{\mathrm{d}t} = Ka(1-a)$$

Differential equation of SI-model

Where:

a is the proportion of vulnerable machines which have been compromised.

K is the rate of infection and K is just assumed constant = 1.8 and does not depend on *"processor speed, network connection, or location of the* infected machine".

The solution of that differential equation will be as:

$$a = \frac{e^{K(t-T)}}{1 + e^{K(t-T)}}$$

Logistic curve describing the rate of infection

Where: T is a time which fixes when the incident happens.

And according to [19]spreading of infections has two phases at start begun quickly which refers by epidemic phase then slows down and refers by steady state. And SI-model could be used to study infection in the network where population is divided to(S)of computer nodes they are considers as vulnerable nodes and (I) of infective which can pass the infection to the susceptible nodes.

Spreading of computer worms can be described [19] by differential equation as:

$$\frac{\mathrm{d_i}t}{\mathrm{d}t} = \beta i(t)(N - i(t))$$

Differential equation of SI-model

Where N is the number of susceptible nodes for all t and $N(t) = i(t) + s(t)$. In this equation the rate of passing infection β is just assumption and constant also.

SIS model

SIS model has ability to stop the information spreading before all individuals become infected [12]. If some node recognizes that there are many infected among its last communication with neighbors, is not going to pass the disease because it became old. And in this model the removed individuals can get infection again.

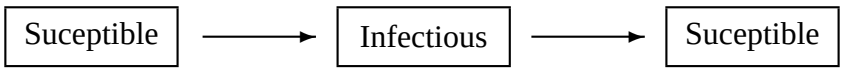


Figure 3.3: SIS model

This model [15] also is one of the simplest models of network epidemic models. This model consists of just two states, the susceptible (S) and the infectious (I) states. Susceptible node can be infected by some infected neighbor at some time step

and become infectious. During the same time step the infectious nodes will be exposed to cure by some probability and become susceptible again. And due to that [20] nodes will change their stated from susceptible to infectious and vice versa many times.

This model [15] does not consider removal properties (death, immunization or protection).In the SIS model [4] there are rate of infection and a recovery, i.e. individuals can be susceptible again.

This model is uses to study the worm propagation while some nodes are out of work for a short time but still there are not cured from infections, for instance when the some infected computer is turned off for some time.

SIS model can be described by the differential equation:

$$\frac{d_i(t)}{\mathrm{d}t} = \beta\bar{d}(1 - i(t))i(t) - \gamma i(t) \tag{3.5}$$

Total rate of infection and recovery of nodes at time t

Where:

- β is the rate of infection.
- $\bar{d}$ is the average degree of an infections node.
- γ is the rate of recovery.

Recovery of the individuals is proportional to the number of infectious nodes and the rate of recovery (γ).

The solution to Eq.(3.5) is:

$$i(t) = \frac{(1 - \delta)i(0)}{i(0) + (1 - \delta - i(0))e^{-(\beta' - \gamma)t}} \tag{3.6}$$

Logistic curve describing the rate of infection and recovery

Where:

$\beta' = \beta\bar{d}$ and δ is the rate of cure.

Similar to SI model if we have complete graph with n nodes then $\bar{d} = (n-1)$ and the fraction of infected individuals will get this solution:

$$i(t) = \frac{(1-\delta)i(0)}{i(0) + (1-\delta-i(0))e^{-(\beta(n-1)-\gamma)t}} \tag{3.7}$$

Logistic curve describing the rate of infection
and recovery for complete graph

SIR model

SIR model has three states that depending on their status [8]:

1. *"Susceptible (S):"* in this phase the individuals are free for any disease but they can be infected at any time.
2. *"Infectious (I):"* in this phase the individuals have the disease and they can infect others.
3. *"Recovered (R):"* in this phase the individuals have been cured and can not infect any others.

Susceptible individuals will be infected with a constant probability per unit time by infectious individuals whom have contact to them. After they have been infected they will remain for some time before they can be recovered.

SIR model [9] has not latent period, what this means is that individuals are infected as soon as they get contact with other infected nodes.
This model [20] is the extension of the SI model and takes in count the removal state in addition to susceptible and infectious states. In this model, a node can be infected just once.

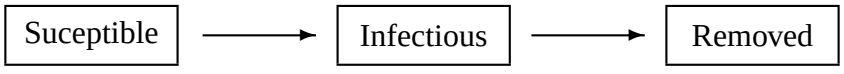


Figure 3.4: SIR model

When the infected node recovers from the disease will get some immunity which can be removed and has not ability to pass the infection anymore [12, 4]. Without any regular immunization the spreading will not be slowing down [15]. But if the nodes with highest degree become immune they will have important effect to prevent propagations growth.

This model is uses to study the worm or virus propagation and at the same time to determine the effects of protections technique such as using of anti-virus program, firewall, intrusion detection software or human countermeasure such as software patching and traffic blocking.

SIDR model

This model [4] deals with four states: susceptible, infectious, detected (in this phase the virus has been discovered but it is not active to spread the infection), and removed.

Figure 3.5: SIDR model

This model is used to study the virus throttling which is an automatic mechanism for restraining or slowing the spread of information [4, 21].

The process of this model contains of two phases: the first is appearing virus signature which lead the node to change its state from susceptible to infectious with some rate. The second phase is detecting the virus. The nodes will be divided in to two classes namely *"throttled and un-throttled"*. If some node belongs to the throttled class and become infected, the infection will not pass through to other nodes and at once will change its state to detected state.

SIRS model

This model [4] has three states: susceptible, infectious, and removed.

Figure 3.6: SIRS model

Immediately after one node becomes removed it will stay in this state for while and this period called *"vigilance period"*, and then will change its state to susceptible again [4, 20].

3.2 Epidemic networks and centrality

Centrality can be measured by betweenness, and betweenness can be considered as measurement of effect that a node has on the behavior of information propagation within the network [22] . Possibly there is another way to measure the centrality namely the degree of node which indicates the number of contact with other nodes in the network. And the degree represents in some way the popularity of a node. But the most powerful way to measure the centrality of some node is closeness which depends on the shortest possible path between this node and other nodes. Closeness measures the lasting of spreading information from one given node to all other nodes in the network.

Betweenness it seems to be considered as location of a node on the paths between others as we can see from Figure (3.7) where nodes A and B lay between tow groups of nodes they consider as *"bridged"* and they obtain the highest betweenness. Thus they represent the shortest path between any tow nodes from both groups and that indicates A and B have important roles of information flow from sources to targets. But node C obtains the lowest betweenness because none of shortest path goes through it.

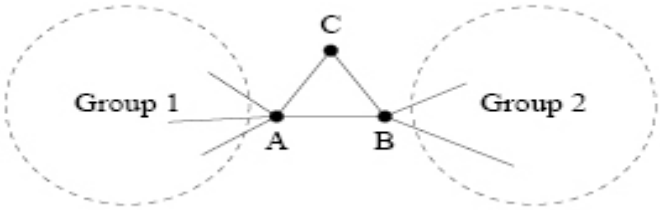


Figure 3.7: Nodes A and B have highest betweenness than node C

The equation of betweenness of node i according [22] is:

$$b_i = \frac{\sum_{s<t} g_i^{(st)}/n_{st}}{\frac{1}{2}n(n-1)} \tag{3.8}$$

Betweenness of node i

Where:

s and t	are two nodes.
n	is the total number of the nodes within the network.
$g_i^{(st)}$	is the number of shortest path between node s and t which pass throw node i.
n_{st}	is the total number of shortest path between s and t.

But since using this method to measure the betweenness is too hard and difficult if we compared it with measuring degree of nodes and since they are too much correlated it is better to use degree as measurement [22, 23]. Nodes with high degree have more influence than nodes with low degree.

Centrality [24] depending on nodes degree,*"ie, its number of neighbors"* and can be related to spreading power. Degree of any node can be as represented in Eq.(3.9)

$$k_i = \sum_{j=nn(i)} \tag{3.9}$$

Equation of nodes degree

Where $nn(i)$ is the nearest neighbors of node i.

Eq.(3.9) is not quite enough to measure the centrality [24] , so by involving neighbors centrality may obtain the smoother measurement of centrality as in Eq.(3.10):

$$e_i = \frac{1}{\lambda} \sum_{j=nn(i)} e_i \tag{3.10}$$

Nodes centrality

And Eq. (3.10) can be rewritten as:

$$Ae = \lambda e \tag{3.11}$$

Where:
A is an adjacency matrix
e is a vector of centrality

Note: Eq. (3.11) is equivalent with Eq.(2.3).

And as we mentioned under section (2.4 Principal eigenvector) here e represents principal eigenvector, that insures by taking the largest eigenvalue (λ_{max}).Then eigenvector centrality values are different from node to node in a smooth way that because the most central node surrounded by important nodes too. That distinguishes the measurement by eigenvector centrality from measurement by degree centrality. The reason is that as in [24, 6] even one node has high degree may do not has high centrality because that node has no connection with important nodes. This is mean on other words that eigenvector centrality takes in count the properties of neighborhoods but node degree does not.

This will lead us to deduce that central node is surrounded by many nodes with high eigenvector centrality and isolated node is poor from that property.

The author of [24] assumed that:
"Eigenvector centrality (EVC) is a good measure of spreading power".
If this assumption is true then the isolated nodes (separate nodes from other nodes) will never have large spreading power.

Spreading will take place from node to node and depends on the centrality of the infected node. Neighbors of infected node will be infected and they will infect their own neighbors and so on. If node has high spreading power it will infect other nodes rapidly than other nodes with low spreading power.

If infection starts from nodes with low spreading power then will reach after while to the nodes with higher spreading power and at end reaches the remained nodes with low spreading power. During these stages the most important nodes will be completely infected and at this time the rate of infection will reach maximum and this lead to obtain *"saturation"* and then infection moves again toward nodes with low eigenvector centrality.

From this conclusion one can obtain the curve with S-shape which can be divided in to three parts or stages. According [24, 25] the first one represents the earlier stage of infection which is known by its flat part. That because nodes with low spreading power are infected at this stage and the curve goes *"uphill"* not so quickly. Then follows by second part which represents the *"saturated"* stage that because nodes with high spreading power are infected and the infection will spread rapidly toward top of infection and that called *"top of the mountain"* where infection increases exponentially. Third part of the curve represents the infection of remainder of the nodes with low spreading power and the rate of infection at this time is slow and the infection seems to be almost linearly.

If we look at Figure (3.8) we can easily see those three stages from the S-shaped curves [25]. This figure shows the rate of infection as function of time which is carried out by applying SIR (Susceptible-Infected-Removed) model on a network with some rate of infection and recovery.

A network with multiple region [24] shows also S-shape even each region has own S shape that because the infection curve for the whole network represents the sum of all infection curves which belong to those regions.

In this paper [24] SI (Susceptible, Infected) model be implemented where if a node infected it will remain infected for ever. At $(t = 0)$ initially there is just one infected node then this will goes to infect all other nodes with fixed probability p per unit time.

Figure (3.9) shows the cumulative S-shape of one region with infection probability $(p = 0.05)$ within *"Gnutella"* graph where the most important nodes are infected at the beginning of the time steps.

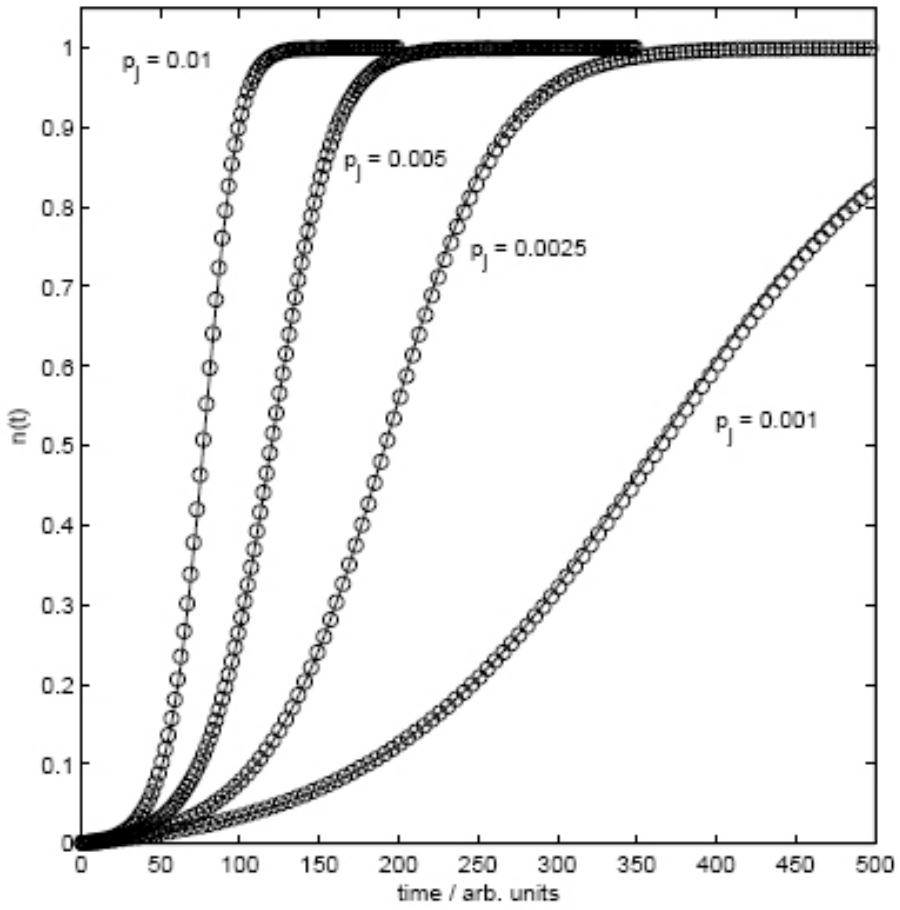


Figure 3.8: Curve of rate of infection (S-Shape)

Figure (3.10) is the same as figure (3.9) in all conditions unless the infection probability is $(p = 0.6)$ and this lead to compressing of time scale for infected nodes. As we can see from figure (3.9) all nodes need 120 unit time to be infected with $(p = 0.05)$ but in figure (3.10) one can see they need almost just 14 unit time steps to be infected when $(p = 0.6)$.

The author of [24] would like to show a side of spreading power as an outline that because they has not get the correct result for their SI spreading process.

First the author was interested to find out an expression for *"infection coefficient"* $C(i, j)$ to be as description for spreading infection from node i to node j and reverse and depends on path from i and j as shown in Eq.(3.12).

$$C(i,j) = \sum_{h=1}^{max} w^h NSR^h(i,j) \tag{3.12}$$

Infection coefficient

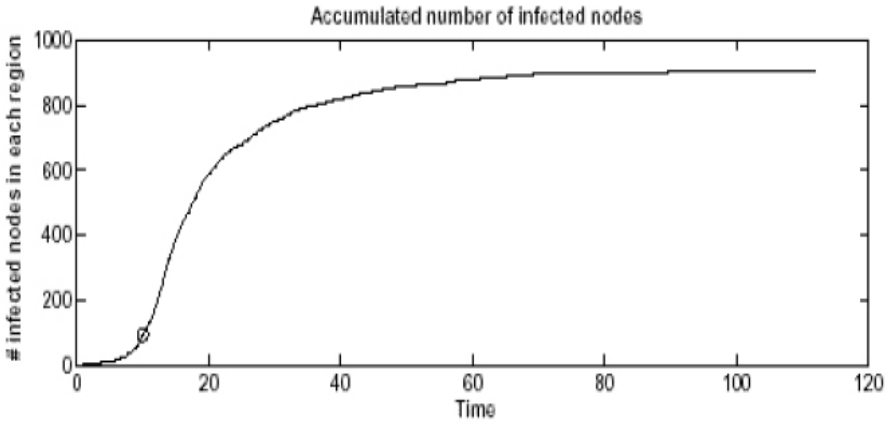


Figure 3.9: Cumulative S-shape with $p = 0.05$

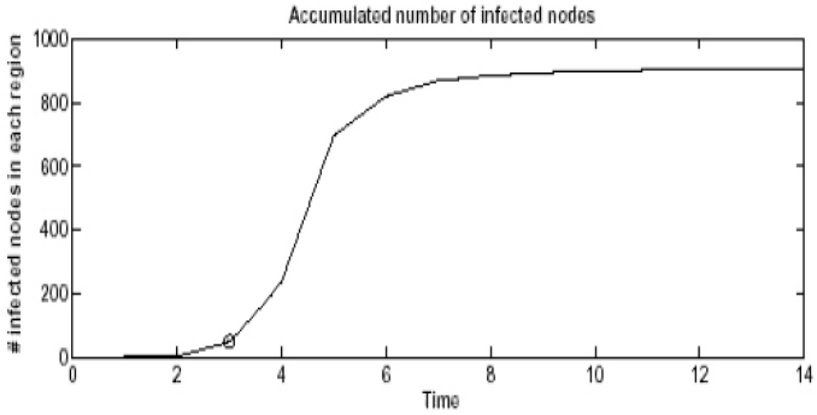


Figure 3.10: Cumulative S-shape with $p = 0.6$

Where:

$C(i,j)$	is infection coefficient
i and j	is tow nodes (i will infect$j$)
h	is path length
NSR	is a an adjacency matrix *"Non-Self-Retracing"* path
w	is some positive weight and $w < 1$

Eq.(3.12) gives the *"infection coefficient"* for each node, and now it easy to apply that equation to fine the spreading power for node i. To do so one must add all infection coefficient $C(i,j)$ which involves node i as shown in Eq.(3.13).

$$S(i) = \sum_{i} C(i,j) = \sum_{i} \sum_{h=1}^{max} w^h NSR^h(i,j) \tag{3.13}$$

Spreading power of node i by excluding Self-Retracing path

Where: Where $S(i)$ is indicate the spreading power.

Here again one can say each node with high infection coefficient has high spreading power.But since NSR^h is still not having a general expression so it will be substituted with as in Eq. (3.14):

$$S^a(i) = \sum_{i} \sum_{h=1} w^h A^h(i,j) \tag{3.14}$$

Spreading power of node i by including Self-Retracing path

Where A is an adjacency matrix.

Well this mathematical theory of the spreading power does not gives the exact solution because of infection maybe doubled when spreading from j to i also taken with.

Chapter 4

Methodology

To approach from satisfied answer to our research question and our predictions; we shall try to exhibit our method to be easier to follow the steps of solving the problem.

4.1 Research Subjects

The subjects of this research are nodes (hosts), which they can be considered as one population. Since we shall test our developed mathematical SI-model for measurement so all nodes in our population (network) are infected or susceptible.

Our network has these properties:

1. The population consist of N nodes.

2. Just one node infected (I) and has ability to infect all other nodes by infection's rate (τ).

3. ($N-1$) nodes are susceptible (S) which they have no idea about the infection but they have ability to get it.

4.2 Research Tools

We shall use several tools for testing and analyzing data.

1. Mathematica:
   Mathematica is a popular tool for computer algebra system and it is a powerful programming language. Mathematica is helpful to: do numerical and symbolical computation, to analyze and visualizing data, to do numerical modelling and simulation, solving simple and complex systems and much more [26].

2. Excel:
   Excel is a powerful program to store information in columns and rows and then can be organized as we want. Excel deals with number and text. Numbers called values and text often called label. Using excel for numerical solution and making and displaying graph from data which stored in the tables. We can use excel for many mathematical and statistical calculating to fine for instance standard deviation, mean, average, max, min, and sum. Furthermore we can choose many types of chart depending on our demand such as: line, bar, pie, cylinder, and cone [27].

3. ORA (Organizational Risk Analyzer):
   The Organizational Risk Analyzer (ORA) is a statistical analysis package for analyzing complex systems as dynamic social networks. By using ORA we can generate and visualize network graph and shows nodes connection with each other by edges.We can easily rotate, isolate, and add nodes as demanded. We can use ORA for sketching many types of chart such as: Bar Chart, Scatter Plot, Histogram, and Heat Map. Furthermore we can use ORA for generating many types of reports for instance, risk, management, immediate impact, and Sphere of influence. We can use it for many other tasks such as comparing organizations to each other, optimizing network structure, and for identify subgroups [28].

4. In addition we used Online Matrix Calculator [1] for calculating eigenvalues and their corresponding eigenvectors.

4.3 Procedures

1. Using ORA to generate two types random network graph (populations); first, is a small network graph and another one is a large network. Small network graph consist of just 12 nodes and large network graph consist of 100 nodes. The purpose of that is to get more overview, to identify patterns or trends and to support the result in some way by comparing them; they may supplement each other.

2. Trying to calculate PEV for each graph.

3. Proportionate to the values of PEV. The purpose of that is to enhance data analysis because it is easier to analysis data if there are relatively distributed and it is more meaningful. Then we will use the relative values of PEV as an assumption for the rate of infection of nodes instead of just assuming any constant number as almost all other models have been done to fit their demands. As we have seen from (section 3.1.1 SI-model Eq (3.3)) that the author [4] assumed that β is the rate of infection and at the same time the author used $\bar{d}$ (average of nodes degree)

[1] http://www.bluebit.gr/matrix-calculator/

to get more sensible result. In our model we used PEV which includes implicit degree, betweenness and centrality at the same time that why we do not need to assume any extra constant number to represents the rate of infection.

4. Trying to develop our mathematical method for SI-model which can describe our demand to show the relationship between the rate of information spreading and centrality.

5. Testing our method to figure out whether it satisfies our demands or not.

6. Tracing infections' movements. This will help us to make a cumulative frequency graph to get better view of infected nodes at each time scale, i.e. collecting data spatiotemporally.

4.4 Proportion

To perform the proportion process as we mentioned under section (4.3) we need to:

1. Calculate PEV.

2. Dividing 1 by the sum of all values which we obtained from PEV and the result is denoted by proportion factor (PF).

3. Multiplying each value of PEV by PF to get the proportion form.

Example:

Let us say we have a network graph consist of 10 nodes as in Figure 4.1:

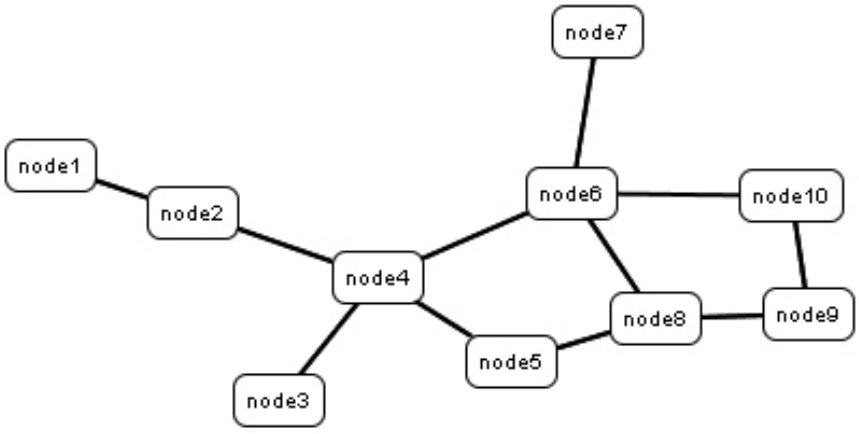


Figure 4.1: Network consist of 10 nodes

The adjacency matrix of Figure (4.1) will be as Eq. (4.1):

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{4.1}$$

Then we used [2] Online Matrix Calculator for calculating eigenvalues and eigenvectors.

Matrix (A) gives eigenvalues $(\lambda) = \{-2.513, 2.513, -1.523, 1.523, -0.764, -0.764, 0.764, 0.764, -0.447, 0.447\}$.

The highest value of eigenvalue is: $\lambda = 2.513$, and the principal eigenvector (PEV) which is associated with this value is as in Eq. (4.2):

$$\vec{v}(A) = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \end{bmatrix} = \begin{bmatrix} 0.090 \\ 0.226 \\ 0.190 \\ 0.478 \\ 0.359 \\ 0.426 \\ 0.170 \\ 0.424 \\ 0.281 \\ 0.281 \end{bmatrix} \tag{4.2}$$

Then next step is to calculate the sum of all values of principal eigenvalue as in Eq. (4.3):

$$\sum_{i=1}^{10} PEV(x_i) = 2.925 \tag{4.3}$$

[2] http://www.linktotheonlinesite.com

Now we shall calculate proportion factor (PF) as in Eq. (4.4):

$$PF = \frac{1}{\sum_{i=1}^{10} PEV(x_i)} = \frac{1}{2.925} = 0.341 \tag{4.4}$$

The final step is to calculate the proportion form of PEV as in Eq. (4.5):

$$PF \times \vec{v}(A) = PF \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \end{bmatrix} = 0.341 \times \begin{bmatrix} 0.090 \\ 0.226 \\ 0.190 \\ 0.478 \\ 0.359 \\ 0.426 \\ 0.170 \\ 0.424 \\ 0.281 \\ 0.281 \end{bmatrix} = \begin{bmatrix} 0.030 \\ 0.077 \\ 0.064 \\ 0.162 \\ 0.122 \\ 0.145 \\ 0.057 \\ 0.144 \\ 0.095 \\ 0.095 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \\ \tau_7 \\ \tau_8 \\ \tau_9 \\ \tau_{10} \end{bmatrix} \tag{4.5}$$

Now after proportion process our $PEV = [0.030, 0.077, 0.064, 0.162, 0.122, 0.145, 0.144, 0.095, 0.095]^T$, and we assume that τ_ith value represent the probability of infection rate for ith node.

4.5 Mathematical method

To develop our mathematical SI-model we would like to bring to mind that our research subjects (see section 4.1) consist of N nodes and just one node is infected (I) and $(N-1)$ nodes are susceptible (S). And we assume that the (τ) is representing the probability for rate of infection after proportionate process; so all susceptible nodes sooner or later will become infected by that probability.

In our method we assume that each node has different ability to pass the infection (information), that according to their location within the network, i.e. nodes have not equal chance to infect others. As mentioned in [8] in the real world the way of contacts between individuals is too far from fully mixed, i.e. they are different in centrality and degree.

To relate degree of the nodes to the rate of infection is not satisfactory, because if two nodes have the same degree will have the same rate of infection too, which is not true. As we can see from Eq.(3.3) the author [4] used β plus $\bar{d}$ (average of degree) to fit or get the suitable result and author of [18] used K see SI-model section (3.1.1) as the rate on infection and this is also just assumed constant and does not depends on network connection or location of the machine.

We used PEV after proportion process to represent the rate of infection that because it includes implicit degree, betweenness, and centrality at the same time. So we have to consider both the degree and location. What we want to say centrality in addition to degree has effect on the rate of passing information.

We shall develop the existing SI model which is considered as one of the simplest epidemic model. Our method depends on differential equation and the idée has come from reviewing previous works see section (3.1.1) [4, 14, 18, 3, 13, 19] with two essential differences between our method and the method of those papers or articles:

1. We have determined that we have just one infected node at the beginning, but [4, 14, 18, 3, 13, 19] do not determined how many nodes infected at the beginning.

2. We shall use PEV (τ) as rate of infection and this value as we mentioned includes implicit nodes centrality and degree, but the rate of infection in [14, 18, 3, 13, 19] depends on just assumed numerical number (β) and in [4] depends on ($\bar{d}$) (average of degree) plus (β) of nodes plus that assumed number.

Since in this model we are depending on just one factor namely the rate of infection (τ) so we do not have any recovery state during the infection process. In this case we suppose that just one node is infected and all other nodes within the network are susceptible, and when they get infected they will remain as they are with out getting cured for ever. We do so because we are interesting to find and determine how one infected node can infect all other nodes within the network depending on principal eigenvector and centrality of that infected node, i.e. depending on the position and degree of that infected node.

The spread of information in this model will depends on the connection between nodes, in other words any infected node will infect all connected susceptible neighbors and each neighboring node will infect its susceptible connected neighbors and so on.

Now we can represent the whole population (nodes) by Eq. (4.6)

$$N = S(0) + I(0) \tag{4.6}$$

As we supposed that we have just one infected node and $(N-1)$ susceptible nodes at start which represented by Eq. (4.7) and Eq.(4.8):

$$I(0) = 1 \tag{4.7}$$

$$S(0) = N - 1 \tag{4.8}$$

So the number of susceptible nodes at time t will be like initial number of susceptible nodes minus newely infected nodes $I(t)$ as in Eq. (4.9):

$$S(t) = S(0) - I(t) \tag{4.9}$$

Substituting S(0) from Eq. (4.8) by S(0) in Eq. (4.9) we get Eq. (4.10) which is represents number of susceptible nodes at time t:

$$S(t) = N - I(t) - 1 \tag{4.10}$$

since the number of infected nodes will change (increase) by each unit time, so it is suitable to use a differential equation as in Eq.(4.11) to represent this process and the change factor is (τ) infection rate.

$$\frac{dI}{dt} = \tau I(t) S(t) \tag{4.11}$$

By setting Eq.(4.10) in Eq.(4.11) we get:

$$\frac{dI}{dt} = \tau I(t)(N - I(t) - 1) \tag{4.12}$$

Since Eq.(4.12) is separable, we will separate it as in Eq.(4.13):

$$\frac{1}{I(t)(N - I(t) - 1)} dI = \tau dt \tag{4.13}$$

Eq.(4.13) is difficult to be integrated so we should split up the left side of the equation to be easier for integration as in Eq.(4.14) where A and B are constants.

$$\frac{1}{I(t)(N - I(t) - 1)} = \frac{A}{I(t)} + \frac{B}{(N - I(t) - 1)} \tag{4.14}$$

$$\frac{1}{I(t)(N - I(t) - 1)} = \frac{A(N - I(t) - 1) + BI(t)}{I(t)(N - I(t) - 1)} \tag{4.15}$$

Then from Eq.(4.15) we get $A(N - I(t) - 1) + BI(t) = 1$ as in Eq.(4.16) since denominator for both sides in Eq.(4.15) is equal.

$$A(N - I(t) - 1) + BI(t) = 1 \tag{4.16}$$

To get the value of A we will set $I(t) = 0$ in Eq.(4.16) and we get Eq.(4.17).

$$A(N - 0 - 1) - 0 = 1 \Longrightarrow A(N - 1) = 1 \Longrightarrow A = \frac{1}{(N - 1)} \tag{4.17}$$

Then to get the value of B we will set $(N - I(t) - 1) = 0$ in Eq.(4.16) and we get Eq.(4.18)

$$A * 0 + B(I) = 1 \Longrightarrow B = \frac{1}{(I)} \tag{4.18}$$

and since we set $(N - I(t) - 1) = 0 \Longrightarrow I(t) = (N - 1)$ and by setting this value in Eq.(4.18) we will get Eq.(4.19).

$$B = \frac{1}{(N - 1)} \tag{4.19}$$

Now we have got $A = \frac{1}{(N-1)}$ and $B = \frac{1}{(N-1)}$ and by setting these values in Eq.(4.14) we will get Eg.(4.20).

$$\frac{1}{I(t)(N - I(t) - 1)} = \frac{\frac{1}{(N-1)}}{I(t)} + \frac{\frac{1}{(N-1)}}{(N - I(t) - 1)} \tag{4.20}$$

and Eq.(4.20) gives:

$$\frac{1}{I(t)(N - I(t) - 1)} = \frac{1}{N - 1}\left(\frac{1}{I(t)} + \frac{1}{(N - I(t) - 1)}\right) \tag{4.21}$$

Then we will set Eq.(4.21) in Eq.(4.13) and the result will be as Eq.(4.22):

$$\frac{1}{N - 1}\left(\frac{1}{I(t)} + \frac{1}{(N - I(t) - 1)}\right) dI = \tau dt \tag{4.22}$$

Now we are going to integrate both sides of Eq.(4.22) and this will give:

$$\frac{1}{N-1}\left(\int \frac{1}{I(t)}dI + \int \frac{1}{(N-I(t)-1)}dI\right) = \tau \int dt \tag{4.23}$$

$$\ln I(t) - \ln(N - I(t) - 1) = \tau(N-1)t + c \tag{4.24}$$

where c is constant

$$\ln\left(\frac{I(t)}{(N-I(t)-1)}\right) = \tau(N-1)t + c \tag{4.25}$$

$$e^{\ln\left(\frac{I(t)}{(N-I(t)-1)}\right)} = e^{\tau(N-1)t+c} \tag{4.26}$$

$$\left(\frac{I(t)}{(N-I(t)-1)}\right) = e^{\tau(N-1)+c} = e^{\tau(N-1)t}e^{c} \tag{4.27}$$

By setting $e^c = C$, where C is constant then we will get:

$$\left(\frac{I(t)}{(N-I(t)-1)}\right) = Ce^{\tau(N-1)t} \tag{4.28}$$

Initializing:

$$\text{when} \qquad t = 0 \Longrightarrow e^{\tau(N-1)*0} = 1 \qquad \text{and} \qquad I(0) = 1 \tag{4.29}$$

By setting Eq.(4.29) in Eq.(4.28) we will get Eq.(4.30)

$$C = \frac{1}{(N-2)} \tag{4.30}$$

And by setting Eq(4.30) in Eq.(28) we will get Eq.(4.31)

$$\left(\frac{I(t)}{(N-I(t)-1)}\right) = \frac{1}{(N-2)}e^{\tau(N-1)t} \tag{4.31}$$

Now to get the final equation of logistic curve of rate of infection we just continue to do some algebraical steps.

$$I(t) = \left(\frac{(N - I(t) - 1)}{(N-2)}\right) e^{\tau(N-1)t}$$

$$I(t) = \left(\frac{(N-1)}{(N-2)} - \frac{I(t)}{(N-2)}\right) e^{\tau(N-1)t}$$

$$I(t) + \frac{I(t)}{(N-2)} e^{\tau(N-1)t} = \left(\frac{(N-1)}{(N-2)}\right) e^{\tau(N-1)t}$$

$$I(t)\left(1 + \frac{1}{(N-2)} e^{\tau(N-1)t}\right) = \left(\frac{(N-1)}{(N-2)}\right) e^{\tau(N-1)t}$$

$$I(t)\left(\frac{(N-2) + e^{\tau(N-1)t}}{(N-2)}\right) = \left(\frac{(N-1)}{(N-2)}\right) e^{\tau(N-1)t}$$

$$I(t) = \frac{(N-1)e^{\tau(N-1)t}}{(N-2) + e^{\tau(N-1)t}}$$

$$\boldsymbol{I(t) = \frac{(N-1)}{1 + (N-2)e^{-\tau(N-1)t}}} \tag{4.32}$$

Eq.(4.32) represents the logistic curve describing the rate of infection which we will use it to measure the number of infected nodes as a function of time t.

4.6 Graphic Representations of Data

Graphic representation is very useful for data analysis and for describing data which is often difficult to be analysis by just tables or other data forms. We will use polygon and histogram for analysis purpose and to get overview on our data. Polygon is a good representation for data when we need it to be described graphical as a line chart and histogram is a bar chart which represents frequency distribution of data.

4.6.1 Tracing infections' movements

Since one node can not be infected by another node if there is no contact between them, so we shall try to trace infections' movements from node to node and observing the number of nodes will be infected within each unit time. The number of infected nodes within one unit time depends on the number of susceptible neighbors for infected nodes. It means we will collect data spatiotemporally; in other words collecting data at a specific location and at specific time.

Note that by contact we mean for example if a machine is already infected by e-mail virus; this virus shall spread to all who have e-mail address in your e-mail address list regardless to where they are and they become infected too.

4.6.2 Cumulative Representation

The term cumulative means adding the number of frequencies after each unit time. Thus after observing the number of infected nodes at each unit time as in section (4.6.1) we shall add them and use *"ogive"*which is used for graphical representation for cumulative values to show the number of infected nodes as a function of time. This will give us a clear view of infection's curve and can be a good support to our mathematical method which is depends on deferential equation.

Chapter 5

Results and Analysis

5.1 Result from our small network

Let us say we have a small network which consists of 12 nodes as in Figure (5.1):

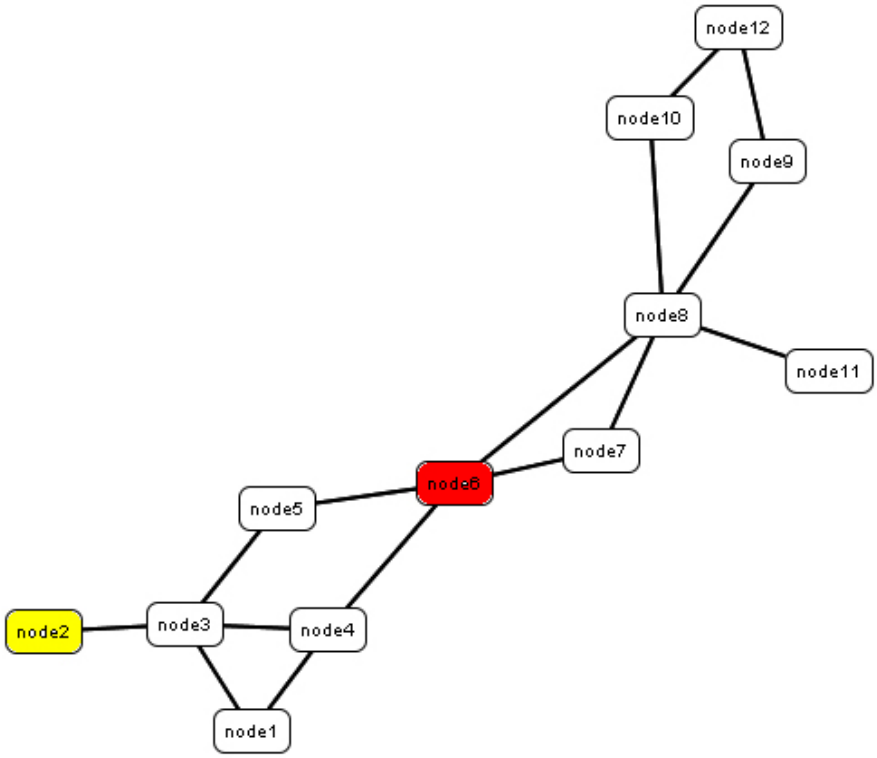


Figure 5.1: Small network consists of 12 nodes

Output from ORA report generator for small network

Node	Betweenness	Eigenvector	Degree
node1	0.0000	0.0709	0.0.1818
node2	0.0000	0.0343	0.0909
node3	0.2091	0.1009	0.3636
node4	0.2545	0.1081	0.2727
node5	0.1273	0.0840	0.1818
node6	0.5545	0.1465*	0.3636
node7	0.0000	0.0979	0.1818
node8	0.5727*	0.1418	0.4545*
node9	0.0818	0.0625	0.1818
node10	0.0818	0.0625	0.1818
node11	0.0000	0.0481	0.0.0909
node12	0.0091	0.0425	0.1818

Table 5.1: Betweenness, Eigenvector, and Degree output from ORA Risk Report

The result of PEV which we obtained from ORA Risk Report is:

```
{
 node1    0.0709
 node2    0.0343
 node3    0.1009
 node4    0.1081
 node5    0.0840
 node6    0.1465
 node7    0.0979
 node8    0.1418
 node9    0.0625
 node10   0.0625
 node11   0.0481
 node12   0.0425
}
```

now we are ranking all nodes from most important nodes to less important nodes as follow:

{

1	Node6	0.1465
2	Node8	0.1418
3	Node7	0.0979
4	Node4	0.1081
5	Node3	0.1009
6	Node5	0.0840
7	Node1	0.0709
8	Node9	0.0625
9	Node10	0.0625
10	Node11	0.0481
11	Node12	0.0425
12	Node2	0.0343

}

Node6 has highest value that indicates node6 ($\tau = 0.1465$) is most important node in this graph and has most centrality in compare with other nodes, and node2 ($\tau = 0.0343$) is considers as the most unimportant node in this graph.

5.1.1 Mathematical Method applied on small network

Now we shall apply our developed mathematical SI-model Eq. (4.32) on our small network which consists of 12 nodes:

$$\boldsymbol{I(t) = \frac{(N-1)}{1+(N-2)e^{-\tau(N-1)t}}}$$

This equation represents the logistic curve for number of infected nodes as a function of time t.

Where $N = 12$ and τ depends on PEV for each node after proportion process.

Note: we have talked about proportion process in section (4.3) and section (4.4) to make analysis easier and useful but in some package such as ORA we do notwant to be worry abut that because the proportion is given automatically..

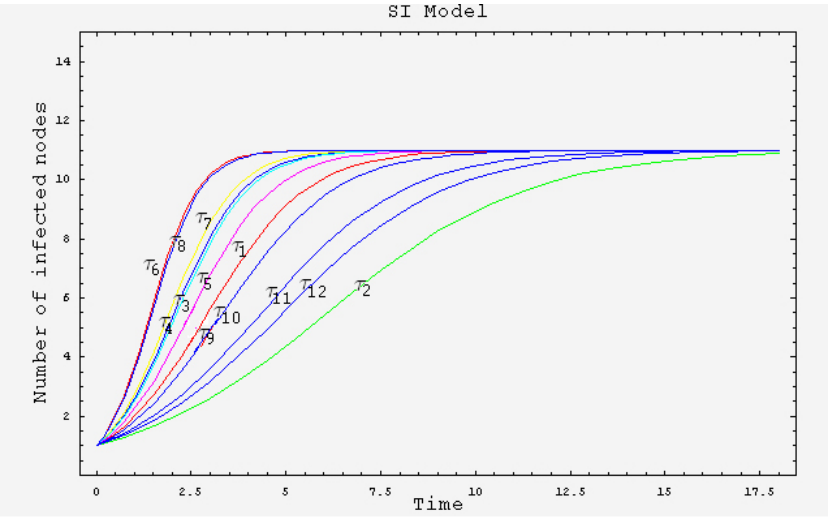


Figure 5.2: screenshot from application of our developed SI-model for all $nodes$ in graph of Figure 5.1

From this figure (Figure 5.2) one can notice the clear S-shape which starts with slow beginning then growth exponential at next stage and finally the curve will take off until all nodes will be infected. Furthermore one can see that when $node6$ which is most important node is infected the rate of infecting other nodes is faster than any other nodes and the curve (rate of infection) growth rapidly.

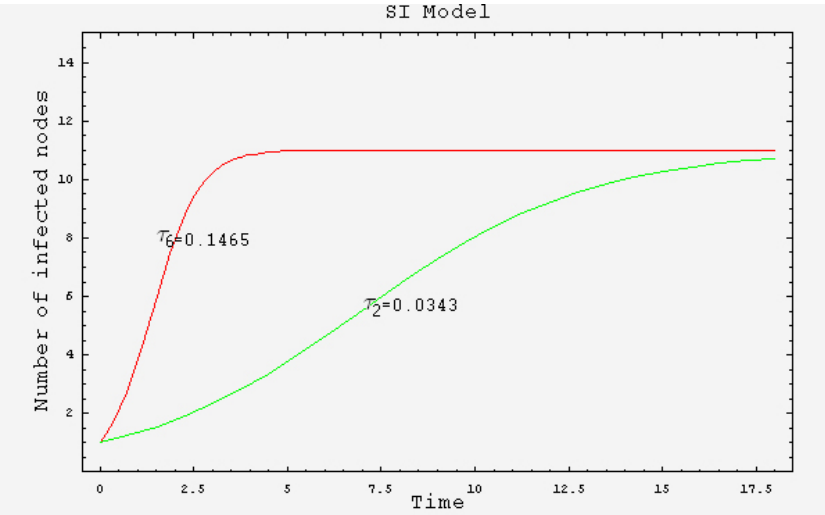


Figure 5.3: Screenshot from application of our developed SI-model for $node_2$ and $node_6$

Figure 5.3 represents the rate of infections by *node2* with $\tau = 0.0343$ and $node6$ with $\tau = 0.1465$. As from ranking nodes we found that $node6$ is most important node and has most centrality and we can see here also from this graph the different between the rate of infection by $node2$ and $node6$. The time scale for infecting all other nodes when $node6$ is infected is shrunk too much in compare to time scale for most unimportant node ($node2$). This indicates that centrality in form of degree and position has crucial role of infection rate.

5.1.2 Tracing infections' movements for small network

Each infected node attempts to infect its susceptible neighbors at rate (β) [20, 29, 30, 4, 31]. And according to [32] *"Propagating viruses: A node in the "susceptible" state will change to "infected" state with the probability (α) only if one of its neighbors is infected, where (α) is the birth rate of the computer virus"*.

Now let us suppose that $node2$ in Figure (5.1) is infected first, and according to principle of each infected node will infect it's nearest susceptible neighbors, $node2$ will infects its susceptible neighbors by some rate and those shall infect their own susceptible (uninfected) neighbors and so on. And we will do the same with $node6$ in Figure (5.1) to compare their ability to spread infection to other nodes in our small graph.

Since network structure has a crucial role for infections spreading [8] and according to [24] *"Since spreading takes place over the links of a network, it is clear that the topology of the network can have a profound influence on the spreading process."*

So if we take a look at Figure 5.4 and Figure 5.6we can see how $node2$ and $node6$ step by step infects all other nodes and infections' movements depending on structure of our small network Figure 5.1.

From Figure 5.1we can see that infections' movements depend on network's structure.

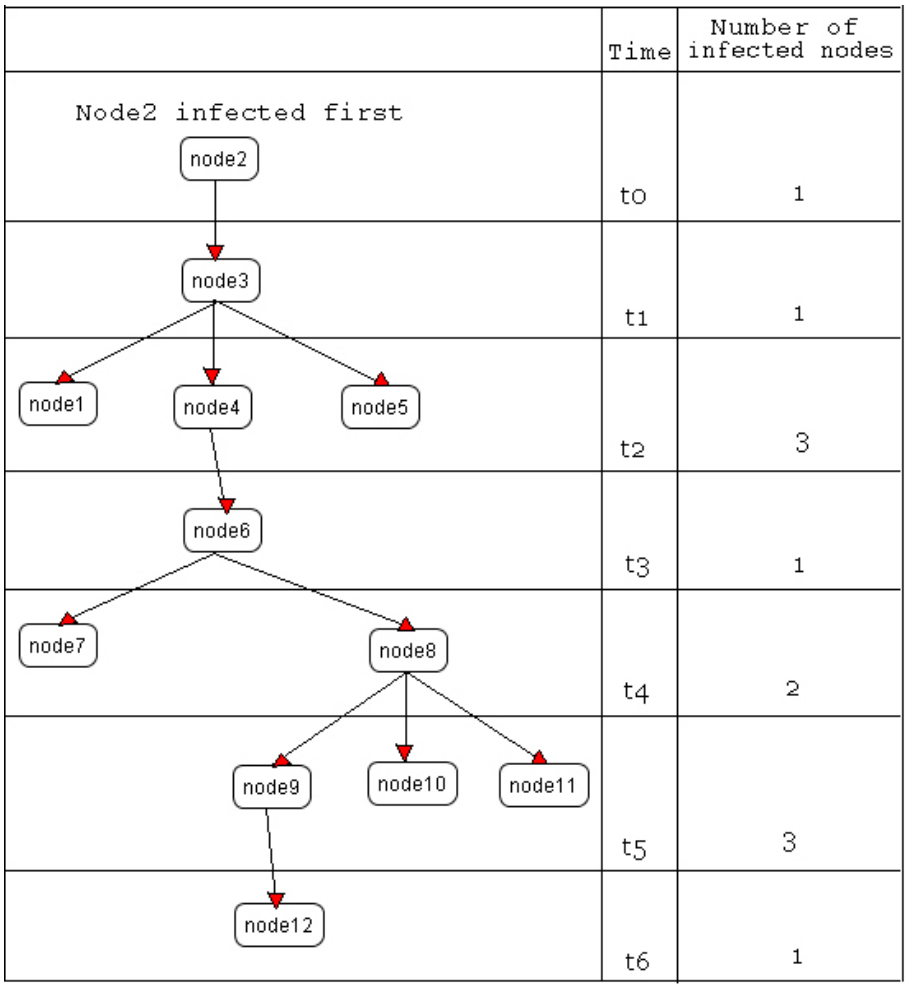


Figure 5.4: Tracing infections' movements from $node2$

From Figure 5.4 we obtained Table (5.2) by tracing infections' movements between nodes.

Table (5.2) consist of two columns as follow:

Frequency column represents the number of infected nodes at each unit time. For example at $t0$ we have just $node2$ is infected and all other nodes are susceptible. At time $t1$ $node2$ will infect only its susceptible neighboring $node3$; which means that we have just one node that is infected at $t1$. Next at time $t2$ $node3$ will infect its susceptible neighbors $node1$, $node4$ and $node5$; it means infecting 3 nodes.Then at time $t3$ $node6$ is the only susceptible node which will be infected by $node4$ or $node5$ because each node if it becomes infected it will remain infected for ever in our SI-model and will not be infected again by other node. Then other nodes by the same way will be infected.

And cumulative column represents the sum of all infected nodes at each unit time.

Cumulative distribution shows the sum of all infected nodes at each unit time in contrast to Frequency distribution which shows just the number of infected nodes at one specific unit time.

Time	Frecquency	Cumulative
t_0	1	1
t_1	1	2
t_2	3	5
t_3	1	6
t_4	2	8
t_5	3	11
t_6	1	12

Table 5.2: Frequency and Cumulative table for infecting other nodes by $node2$ at t_i

The Frequency histogram and cumulative polygon Figure 5.5 represents the frequency and cumulative distribution of Table (5.2). This chart helps us to see clearly how the nodes will be infected at each unit time. Where the horizontal axis represents the time scale for infection and the vertical axis represents the number of infected nodes. From Figure 5.5, we note that, when $node6$ the most important node (the red ball at time $t3$) gets infected the curve of infection will rise up clearly and continues until all other nodes will be infected.

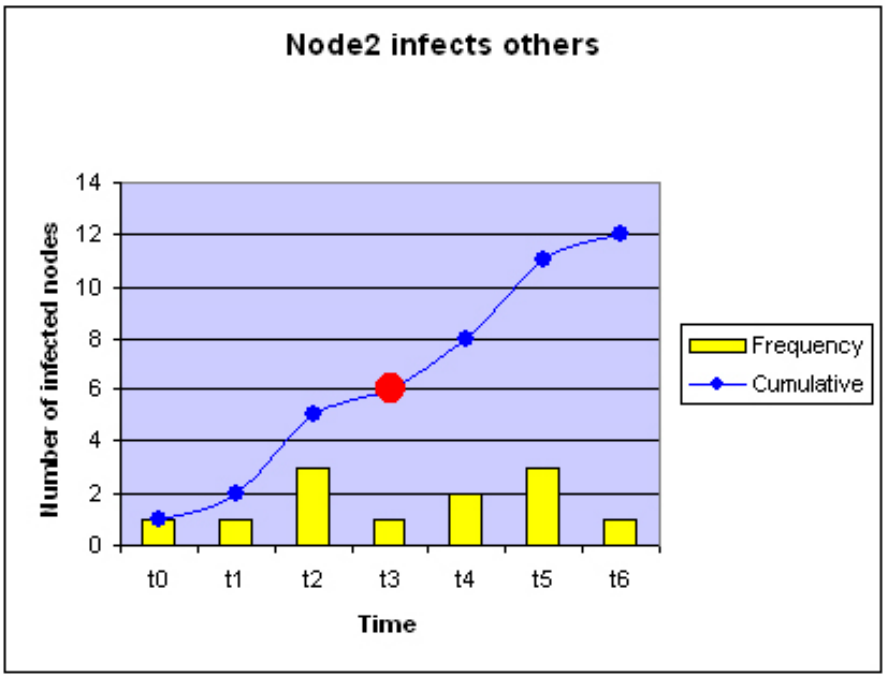


Figure 5.5: The Frequency and Cumulative graphical representation for Table (5.2)

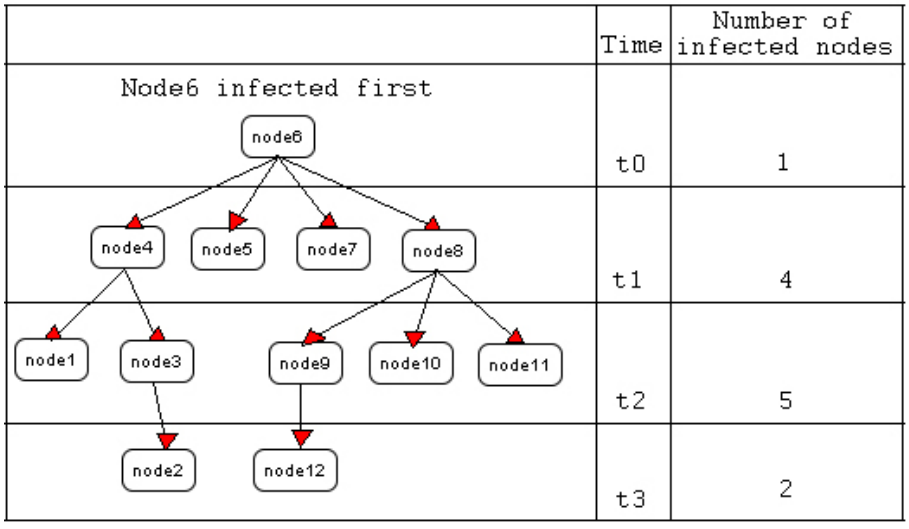


Figure 5.6: Tracing infections' movements from $node6$

Figure 5.6 shows the infections' movements step by step when $node6$ infected first then passing the infection to other nodes. Also here the number of infected nodes at each time scale depends on structure of the network see Figure 5.1.

By the same way as we did when we obtained Table (5.2) we will trace infections' movements from $node6$ and collect the data as in Table (5.3)

Time	Frecquency	Cumulative
t_0	1	1
t_1	4	5
t_2	5	10
t_3	2	12

Table 5.3: Frequency and Cumulative table for infecting other nodes by $node6$ at t_i

The Frequency histogram and cumulative polygon Figure 5.7 represents the frequency and cumulative distribution of Table (5.3). From Figure 5.7, we note that, how the curve increased rapidly when $node6$ the most important node (the red ball at time $t0$) gets infected.

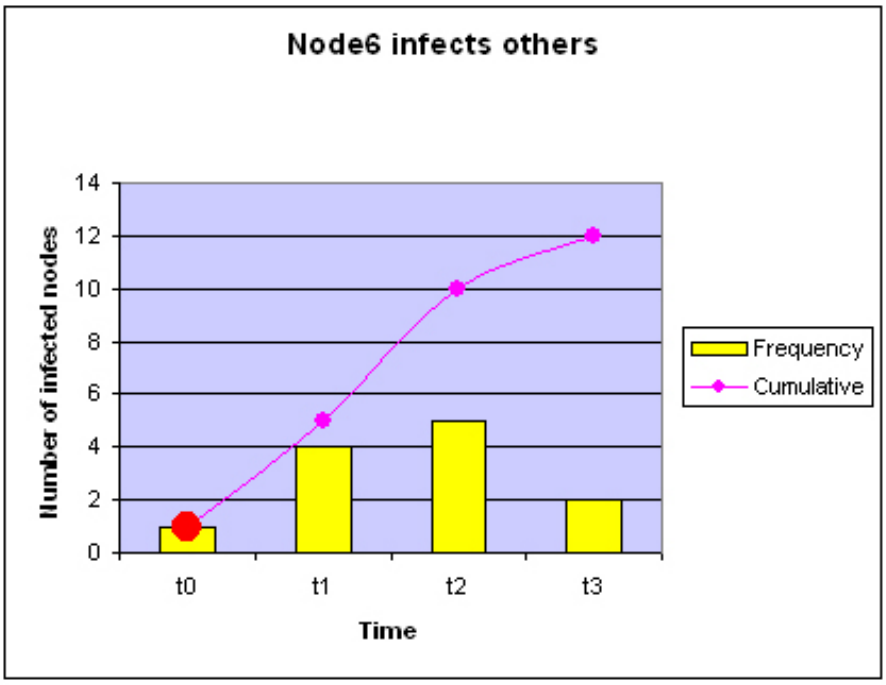


Figure 5.7: The Frequency and Cumulative graphical representation for Table (5.3)

Now we would like to set infections' rates which we obtained from cumulative distribution for both $node2$ and $node6$ side by side in one graph to compare them and see how they have different power to infect other nodes.

Time	Node2 infects other nodes at t_i	Node6 infects other nodes at t_i
t_0	1	1
t_1	2	5
t_2	5	10
t_3	6	12
t_4	8	
t_5	11	
t_6	12	

Table 5.4: Cumulative table for infecting other nodes by $node2$ and $node6$ for our small network (12 nodes)

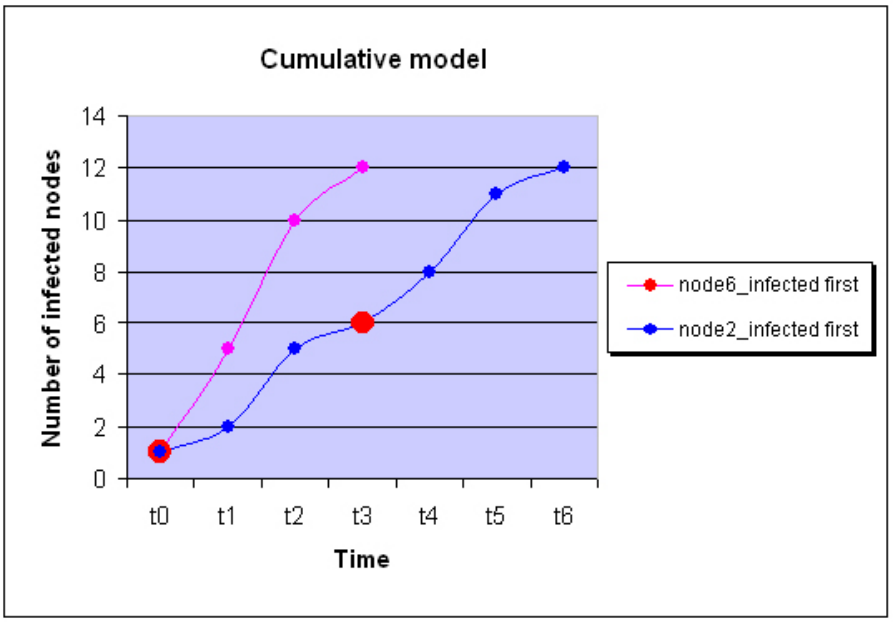


Figure 5.8: The cumulative graphical representation for Table (5.4)

From Figure5.8 one can notice clearly the difference between the rate of infection when $node2$ and $node6$ are infected first, where $node6$ infects other nodes rapidly. Note that when $node2$ is infected at start the most important node $node6$ will be infected at time $t3$.

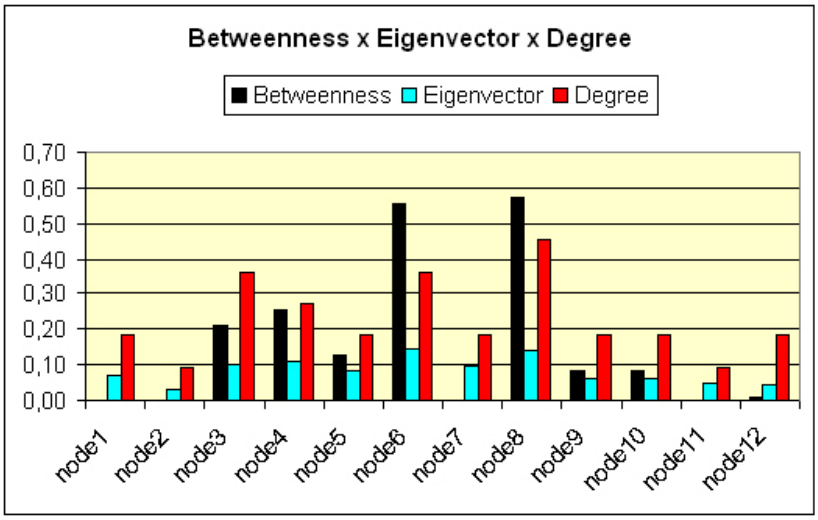


Figure 5.9: Histogram quantities representation of nodes betweenness, eigenvector and degree for Table (5.1)

In Figure 5.9 we used histogram to describe the difference between nodes betweenness, eigenvector and degree. Histogram is a good help to illustrate data and each bar in the graph along x-axis has its height which represents the proportion between nodes with respect to their betweenness, eigenvector, and degree.

If a specific node has highest degree or highest betweenness or both does not mean that that node has most centrality and has most power to infect other nodes. If we look at bar chart for $node8$ this node has highest betweenness and highest degree in compare to $node6$ and also $node3$ has the same degree as $node6$ but $node6$ is most central and has highest PEV. That why $node6$ has highest power of spreading infection.

This means $node6$ is well connected and moreover its connections end almost with well connected nodes too as we can see from Figure 5.1such as $node8$and $node4$. But even $node8$ has highest degree and highest betweenness it connect to not important nodes unless $node6$.

From the graph we can see easily if $node6$ goes down the connections between nodes shall be less than any if any other node will goes down or the amount of damage will be mostor the amount of damage will be most. For example if $node3$ goes down even this node has same degree as node6 still we have good connections.

Thus principal eigenvector is a good measurement for centrality of nodes.

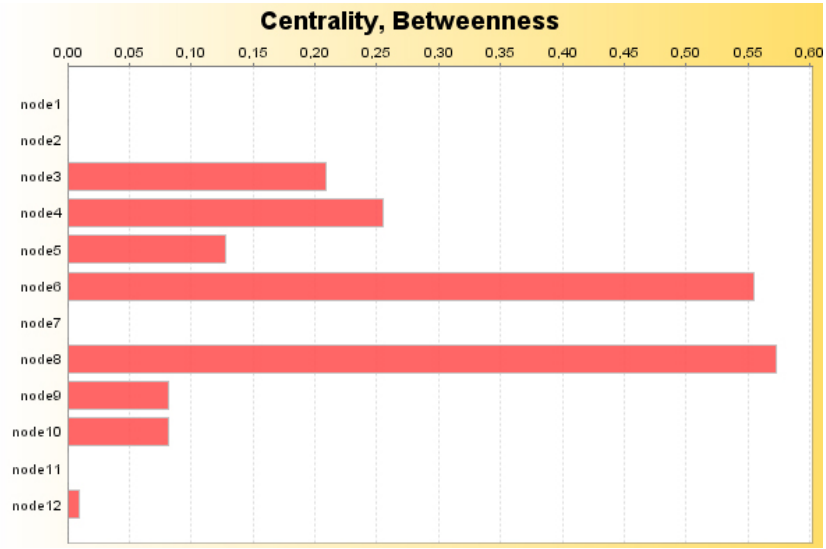


Figure 5.10: Bar chart quantities representation of nodes betweenness

Here in Figure 5.10 node's betweenness expressed by bar chart, where the length of each bar represents the quantities of betweenness. If we observe this bar chart we can easily compare all nodes due to their betweenness. We can see that $node8$ has highest value in contrast to $node1$, $node2$, node7, and $node11$ they have no betweenness at all. If we look at Figure 5.1 it was easy to see that $node2$ and $node11$ has no betweenness but it was not easy for us to see that $node1$ and $node7$ has no betweenness. So this bar chart was a good help to illustrate what was not easy for us to see. From this chart we found that betweenness of nodes does not depend on node's degree, as we can see from Figure 5.1 that the degree of $node4$ $k = 3$ and the degree of $node3$ $k = 4$ but if we look at Figure 5.10 we can notice that $node4$ has higher betweenness than $node3$, that because $node4$ has contact with two important nodes namely ($node3$ and $node6$) but $node3$ has contact with just one important node that is $node4$.

The most powerful nodes in this graph are $node8$ and then $node6$ with respect to their betweenness; if one of them goes down the whole network will be infected. On the other side if one or more than one of $node1$, $node2$, $node7$, and $node11$ get down no thing happens to the whole network. Thus centrality of nodes has crucial role to infect the whole network.

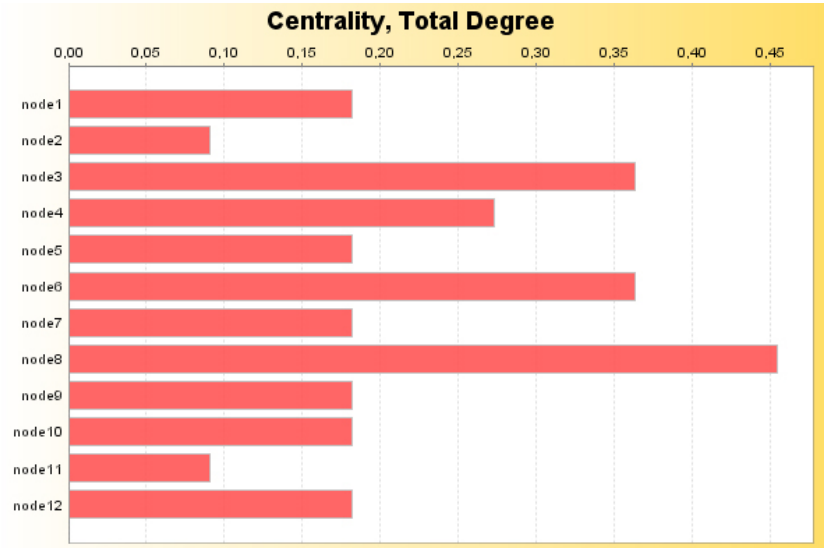


Figure 5.11: Bar chart quantities representation of nodes degree

From Figure 5.11 we can observe that $node8$ has longest bar, i.e highest degree and we can see that $node3$ and $node6$ has the same length, where the length of each bar represents the quantities of degree. But notwithstanding $node6$ has more power to spread the infection because of it position or because of its centrality.

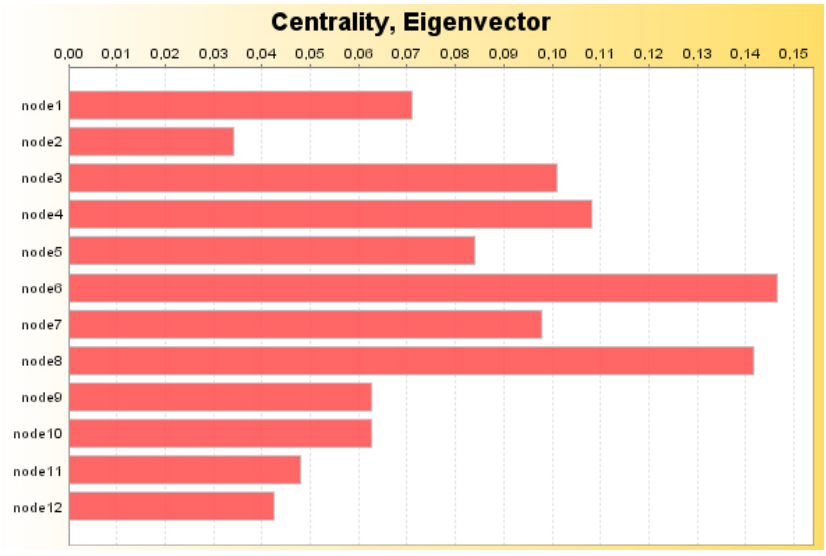


Figure 5.12: Bar chart quantities representation of nodes eigenvector

As we mentioned above since $node6$ has the central position because of its degree and its important connections so from Figure 5.12 we can see that $node6$ has the longest value which indicates that $node6$ is most powerful node in this network.

Thus we will repeat our statement which is saying *"principal eigenvector is a good measurement for centrality of nodes"*.

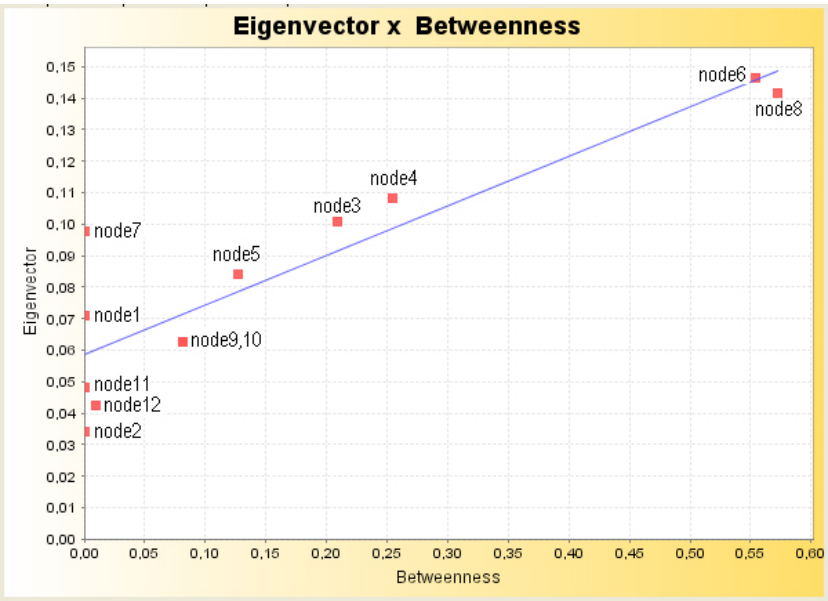


Figure 5.13: Scatter quantities representation of nodes eigenvector versus betweenness.

Scatter chart is a good representation if we have two kinds of variables and we want to know the nature of relationship between them.

From Figure 5.13 we can see that we have independent and dependent variables:

1. Independent variable as betweenness of nodes along x-axis which does not depends on other factors than the position of the node.

2. Dependent variable as eigenvector of nodes along y-axis which depends on other factors such as centrality which depends on degree and betweenness and the position of the node within the network.

When betweenness of node changes the eigenvector of the node changes too, as we can see $node8$ and $node6$ have highest betweenness that effects their eigenvector to be high too but since we have other factors which effects eigenvector so $node6$ is most powerful than $node8$. Note that $node7$ has high eigenvector in some way but that does not change anything from its betweenness where it is zero. That because eigenvector of $node7$ depends on other factors such as degree of $node7$ and the position within this network; where $node7$ has contact to other important nodes such as $node6$ and $node8$ as we can see from Figure 5.1.

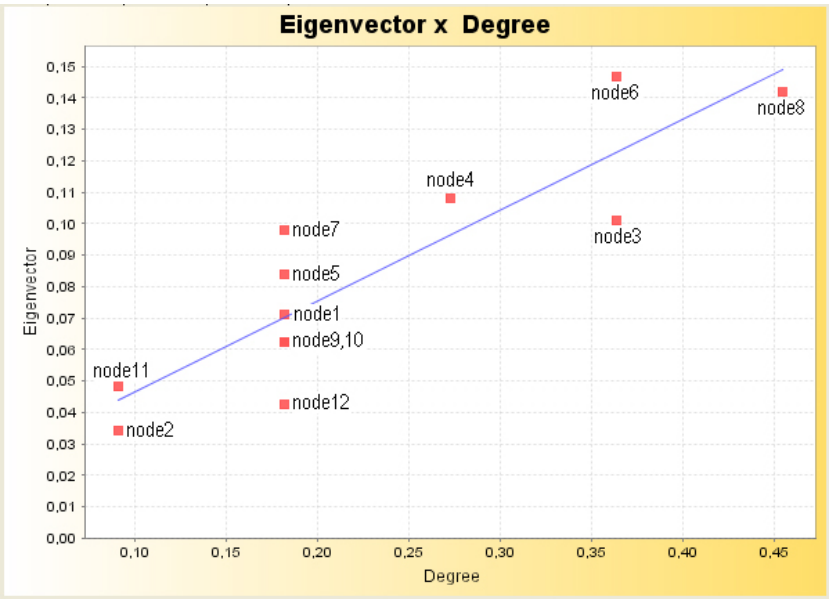


Figure 5.14: Scatter quantities representation of nodes eigenvector versus degree.

Figure 5.14 illustrate the relationship between two variables; independent variable (degree) and dependent variable (eigenvector). Where degree of node does not change (independent) as we can see from Figure 5.14 for nodes 1, 5, 7, 9, 10, and 12 all have the same degree ($k = 2$) but their eigenvector changes (dependent) because of their position (centrality) within our network Figure 5.1. Degree has partial effect on eigenvector of the nodes as we can see from Figure 5.1 by increasing degree eigenvector increase too but sine they are other factors have their effects so we can see $node8$ has highest degree but $node6$ has highest eigenvector which indicate that $node6$ is most important node in our graph Figure 5.1.

5.2 Result from our large network

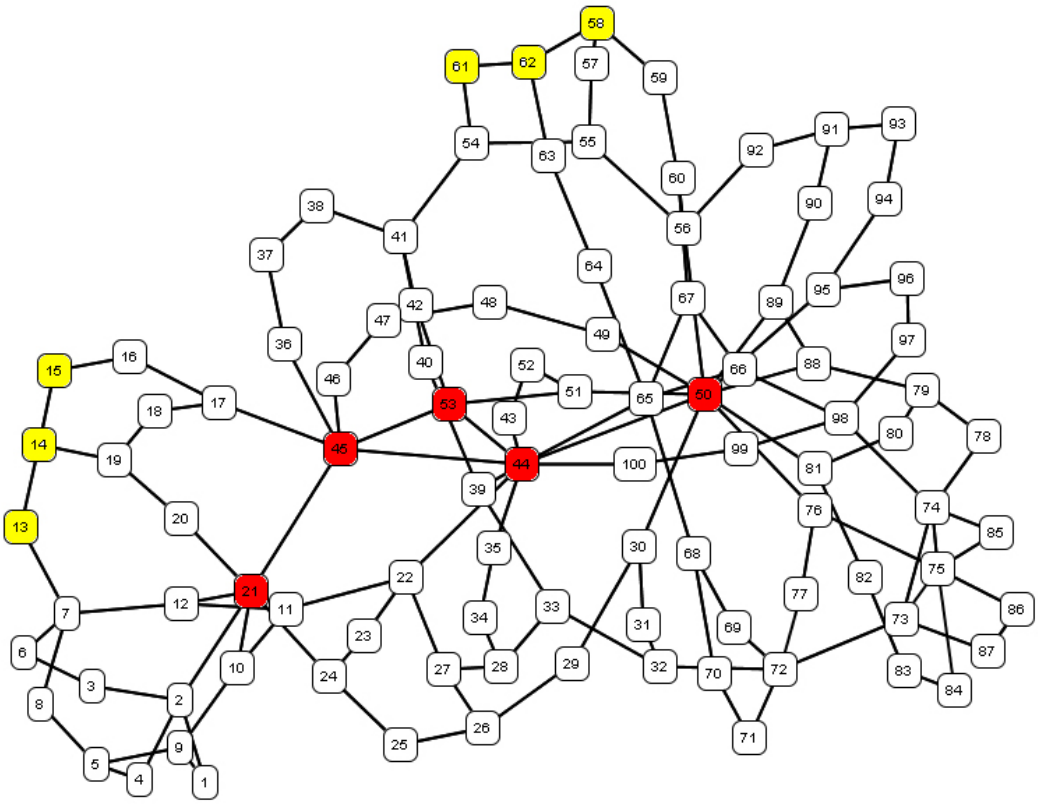


Figure 5.15: Network of 100 nodes

Figure 5.15 represents a screenshot of network which consist of 100 nodes generated by ORA.

From ranking nodes (see Appendix E) we can see that node44 has highest value that indicates $node44$ (τ = 0.0656) is most important node and has most centrality in compare with other nodes then followed by $node50$ and $node45$. And from other side we can see that $node61$, $node62$, and $node15$ are considers as the most unimportant nodes in this graph.

5.2.1 Mathematical Method applied on large network

Now we would like to apply our developed mathematical SI model Eq. (4.28) to show the rate of spreading information (infection) in graph of Figure 5.15 depending on PEV.

$$\boldsymbol{I(t) = \frac{(N-1)}{1+(N-2)e^{-\tau(N-1)t}}}$$

Where N = 100, and τ depends on PEV for each node in Figure 5.15.

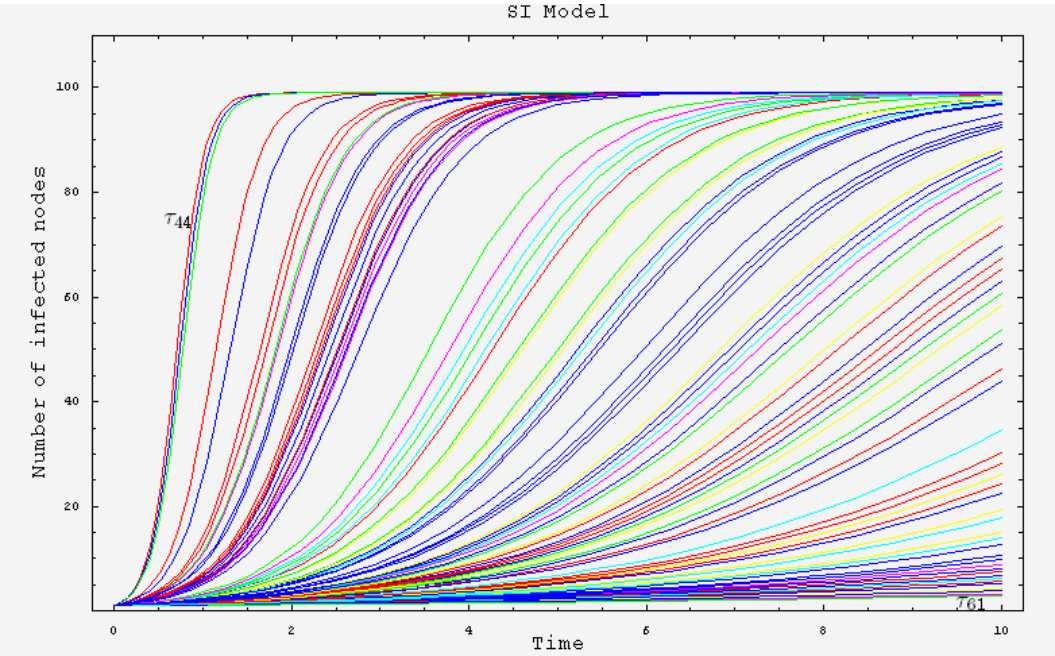


Figure 5.16: Screenshot of mathematical application of our SI model for all nodes in graph of Figure 5.15

From Figure 5.16 we can observe that curves of infections' rates (information spreading) are distributed from highest rate; as we can see when $node44$ with $\tau = 0.0656$ infects all other nodes to lowest rate when $node61$ with $\tau = 0.0011$ infects all other nodes. This indicates that $node44$ is most important node in this network. If we take a look at Figure (C.1, C.2, and C.3) we can see easily that $node44$ has highest betweenness and highest eigenvector and just $node50$ has highest degree than $node44$. This led us to produce that centrality of node does note depend on just degree of node but on position and connection to other important nodes.

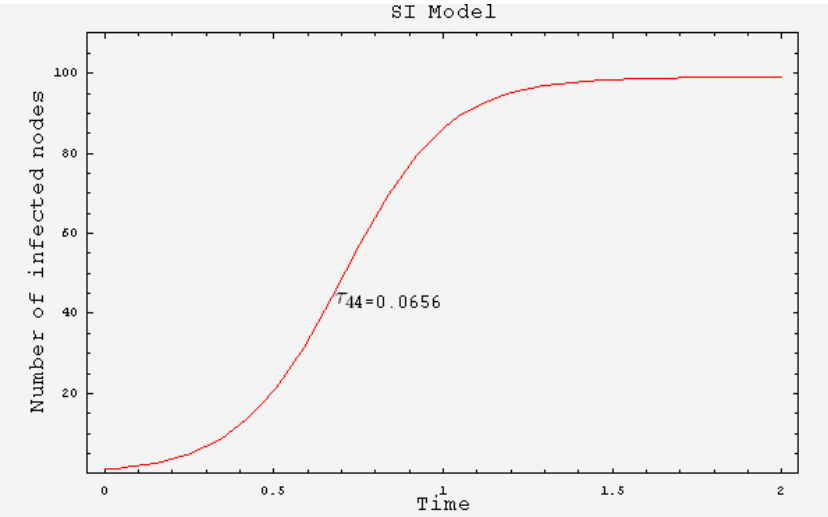


Figure 5.17: Screenshot of mathematical application of our SI model for $node44$

Figure (5.17) represents the rate of infections by $node44$ with $\tau = 0.0656$. Here we can see also the clear S-shape of rate of infection with three stages: slow, exponential and final stage.

5.2.2 Tracing infections' movements for large network

Here again we analyse as we explained in section (5.1.2) according to principle of each infected node will infect it's nearest susceptible neighbors.

By tracing the infections' movements when $node44$ is infected first as in Figure (D.1) (see Appendix D) we have obtained Table (5.5): where column Frequency represents the number of infected nodes at each unit time and column Cumulative represents the sum of all infected nodes at each unit time. The number of infected nodes at each time scale depends on structure of the network see Figure 5.15.

Time	Frecquency	Cumulative
t_0	1	1
t_1	9	10
t_2	26	36
t_3	23	59
t_4	21	80
t_5	15	95
t_6	5	100

Table 5.5: Frequency and Cumulative table for infecting other nodes by $node44$ at t_i

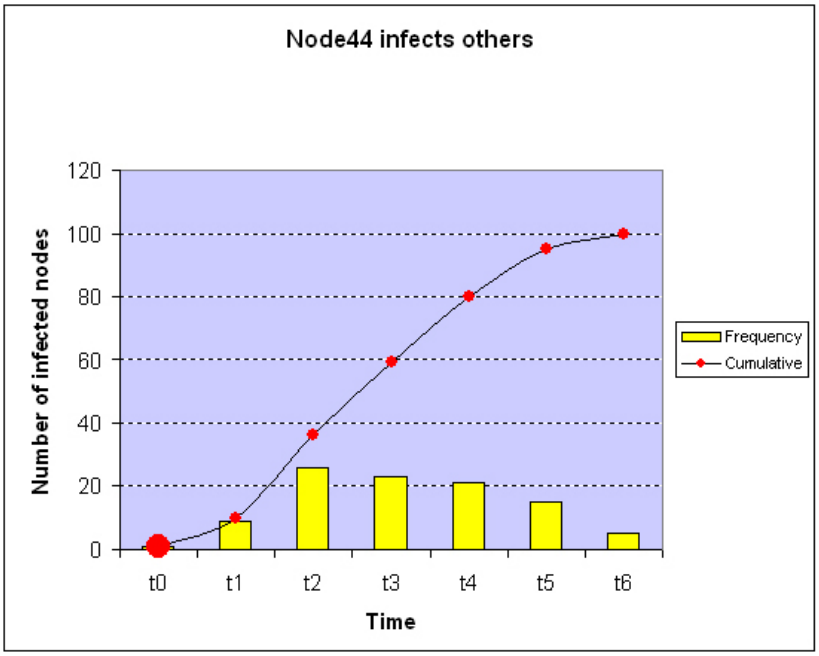


Figure 5.18: The Frequency and Cumulative graphical representation for Table (5.5)

From Figure 5.18 we observe that, how the curve increased rapidly when $node44$ the most important node (the red ball at time t0) get infected and we can compare this with Figure 5.19 to see the different between infections rate at start.

Time	Frecquency	Cumulative
t_0	1	1
t_1	2	3
t_2	4	7
t_3	6	13
t_4	15	28
t_5	22	50
t_6	19	69
t_7	13	82
t_8	10	92
t_9	8	100

Table 5.6: Frequency and Cumulative table for infecting other nodes by $node57$ at t_i

We obtained Table (5.6) by tracing infections' movements Figure D.2 (see Appendix D) step by step when $node57$ infected first then infection spreads to other nodes. Also here the number of infected nodes at each time scale depends on structure of the network see Figure 5.15.

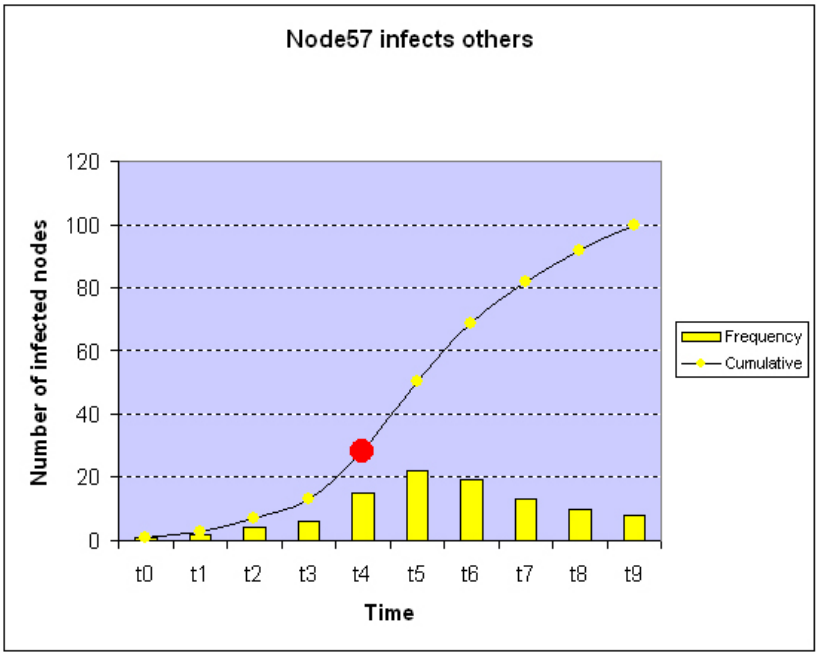


Figure 5.19: The Frequency and Cumulative graphical representation for Table (5.6)

Figure (5.19) represents the frequency and cumulative distribution of Table (5.6). When $node44$ the most important node (the red ball at time $t4$) get infected the curve of infection will rise up clearly. Thus this indicates how $node44$ has its role to spread information rapidly.

Now we would like to compare cumulative distribution as in Table (5.7) for both $node44$ and $node57$ to have a clear view of how different nodes have different power of spreading infection that by plotting them in one graph as in Figure 5.20.

Time	node57$_{infectedfirst}$	node44$_{infectedfirst}$
t_0	1	1
t_1	3	10
t_2	7	36
t_3	13	59
t_4	28	80
t_5	50	95
t_6	69	100
t_7	82	
t_8	92	
t_9	100	

Table 5.7: Cumulative table for infecting other nodes by $node44$ and $node57$ in our large network (100 nodes).

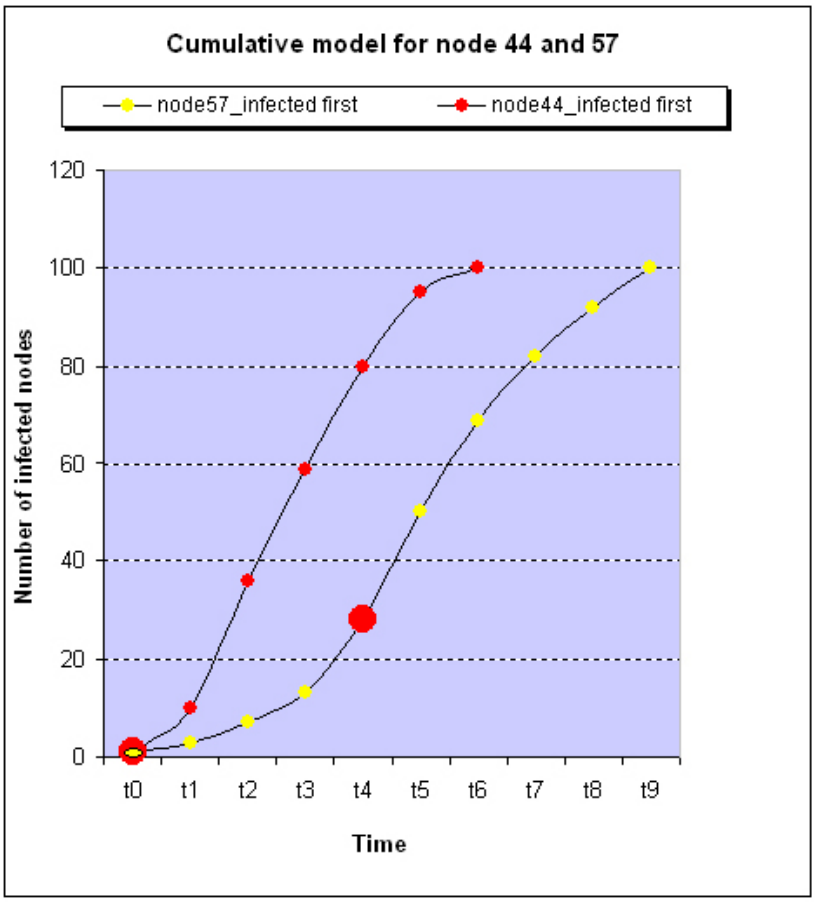


Figure 5.20: The cumulative graphical representation for Table (5.7)

From Figure 5.20 one can notice clearly the different between the rate of infection when $node44$ and $node57$ are infected first. Note that when $node57$ is infected at start the most important node $node44$ will be infected at time $t4$. So nodes with different centrality have different rate of infection. And important node infects other nodes rapidly than unimportant nodes or less important nodes and time of infection shrinks too when important node infects others.

Chapter 6

Conclusions and Discussion

6.1 Conclusions

We investigated during this thesis epidemic network and centrality and the aim was to study the information (infection) spreading, to find most important nodes in the network graph, and finally to answer the research question.

Our research question in section (1.3) was: *"Does principal eigenvector and centrality of the nodes related to the rate of information (infection) spreading?"*

We had two strategies to answer our research question:

- First was depending on our developed mathematical method see section (5.1.1) and section (5.2.1). The result was clear supported our research question as we observed that nodes with highest principal eigenvector (PEV) and most centrality had infected other nodes rapidly than nodes with low (PEV) as we saw from curves of infections see Figure (5.2), (5.3), and (5.16). This is mean principal eigenvector and centrality of the nodes is related to the rate of information (infection) spreading.

- And second depending on tracing infection' movements see section (5.1.2) and section (5.2.2). Here again the answer to our research question was clear as we saw from Figure (5.5), (5.7), (5.18), and (5.19); when most important nodes infected, the curve of infection will increase clearly and rapidly which indicates that these nodes has most centrality and they have highest (PEV) which is mean principal eigenvector and centrality of the nodes is related to the rate of information (infection) spreading.

Furthermore we can observe from Figure (5.8) and (5.20) the different between the rate of infection when one node with high principal eigenvector and high centrality such as $node6$ in Figure (5.8) and $node44$ in Figure (5.20) and one node with low principal eigenvector and low centrality such as $node2$ in Figure (5.8) and $node57$ in Figure

(5.20) are infected first and then infecting other nodes. It was clear that node with high principal eigenvector and high centrality infects other nodes rapidly which is support our research question.

Thus we can see from this conclusion that our hypothesis which states that principal eigenvector and centrality of the nodes is related to the rate of infection in section (1.3) is supported and confirmed too.

And our extra expected prediction in section (1.3) which is stated that *"nodes with different centrality have different curves of infections regardless of their degree average"* is answered quite well. As we can see from Figure (5.14) that for example $node7$ and $node12$ have the same degree ($k = 2$) see also Figure (5.1) but if we look at Figure (5.2) we can observe that $node7$ has different curve of infection than $node12$ that because of their position in the network, i.e. centrality has effect on rate of infection.

The works have been done which concern SI-model see section (3.1.1) [4, 14, 18, 3, 13, 19] is different from our SI-model at rate of infection and at the number of infected nodes at the beginning. When it concerns the rate of infection we have assumed that principal eigenvector PEV (τ) which is represents centrality in addition to degree and betweenness as a rate of infection (information spreading) in contrast to [4, 14, 18, 3, 13, 19] used a numerical assumed constant in addition to degree average see section (3.1.1), or just by a constant probability as in section (3.2) Figure 3.10 and Figure **??**. And when it concerns the number of initially infected nodes our SI-model begins with just one node (initially) as infected but according to [4, 14, 18, 3, 13, 19] the number of infected nodes at the beginning is not determinate.

But the question which SI-model in [4, 14, 18, 3, 13, 19] will face here is if we suppose that one node infected at the beginning and since they depend as we mentioned on a numerical assumed constant which its value is same for all nodes in addition to degree average which lead to may two nodes with same degree average have the same rate of infection which is not true because nodes effect does not depend on just degree average as we explained above.

Finally since we do not know exactly how infection in real world will spread, so it is difficult to be sure about any answer we are getting from our experiment. And us we have noticed from our research almost all previous works depend on assumption to find the rate of infection, i.e. there is no any exact answer because the base is assumption always.

6.2 Results Utilities and Recommendations

Our developed mathematical SI-model Eq.(4.32):

$$\boldsymbol{I(t) = \frac{(N-1)}{1+(N-2)e^{-\tau(N-1)t}}}$$

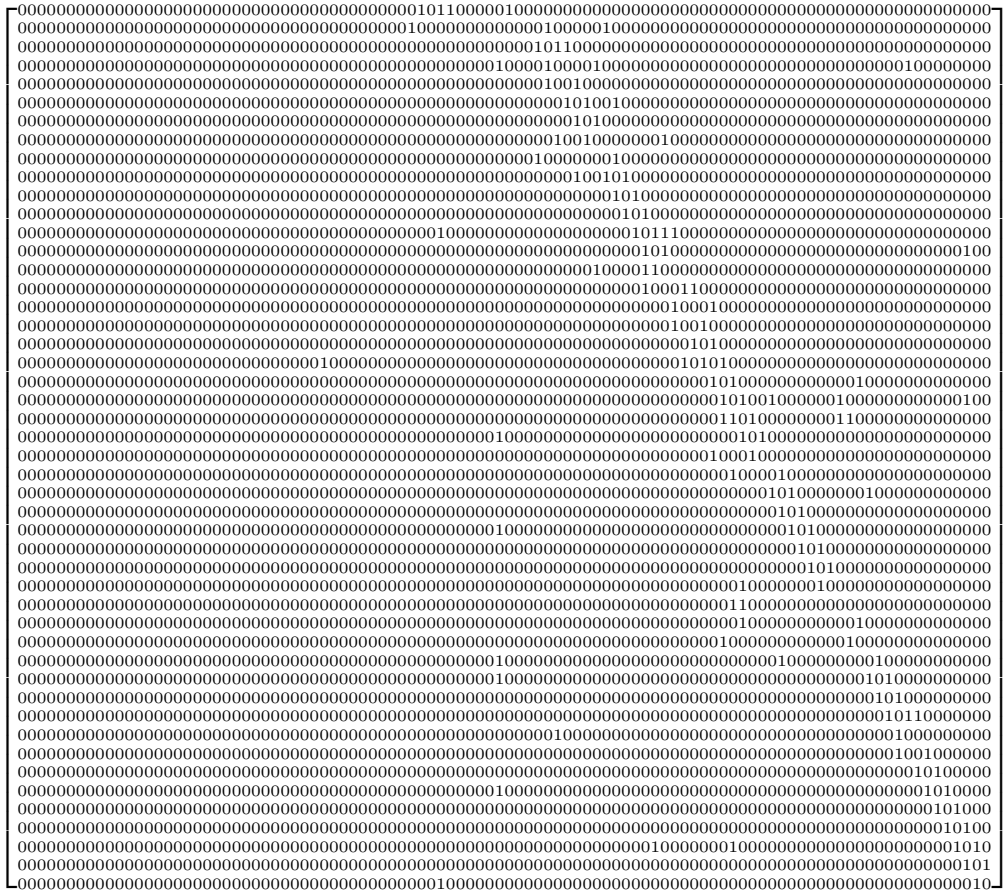
This equation it could be used in any software for network analyzing or simulator or worm (information) propagation programs for describing or showing the logistic curve of spreading power within the network by one specific node. In this case we should take into account that the simulator should calculate the principal eigenvector for chosen node to be as a probability of the rate of infection. For example we have used ORA see section (4.2) this is one of the most user friendly and advanced tool to generate network and use it for analysis, the only thing this software missing from our view is just an epidemic process simulating for nodes of interested network graph. So for example ORA can include our equation into their software program to analyze rate of infections.

Appendix A

Adjacency matrix

The adjacency matrix for network (100 nodes) of Figure 5.15 will be as:

$$A = \begin{bmatrix}
0100000010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
1011000000000000000010000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0100010000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0100100000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0001000110000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0010001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000010100011000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000101000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
1000100001000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000010100000000010000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000001010000000011000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000001000100000000010000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000001000000100000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000001010001000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000101000000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000010100000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000001010000000000000000000000000010000000000000000000000000000000000000000000000000000000 \\
0000000000000000101000000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000100010100000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000001010000000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0100000001110000000100010000000000000000000010000000000000000000000000000000000000000000000000000000 \\
0000000000100000000000100010000000000000000100000000000000000000000000000000000000000000000000000000 \\
0000000000000000000001010000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000010101000000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000010100000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000001010100000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000001000101000000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000010000011000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000100010000000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000101000000000000000000100000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000010100000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000001010000000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000001000100000010000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000001000000100000000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000001000000000100000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000001000000010000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000010100000000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000001000100000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000010000001000100000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000010100000000000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000101010000000000010000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000100000000000100000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000000100000001000000000000000000000000000000000000000000000000 \\
0000000000000000000001000000000000100010001010000100100000000000100000000000000000000000000000000001 \\
0000000000000000100010000000000000010000000101000000100000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000000010100000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000000001010000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000000000101000000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000000000010100000000000000000000000000000000000000000000000000 \\
0000000000000000000000000000010000000000000100001010000100000000000000000001000010000001100000100000 \\
0000000000000000000000000000000000000000000000000101100000000000000000000000000000000000000000000000 \\
0000000000000000000000000000000000000000001000000010000000000000000000000000000000000000000000000000
\end{bmatrix}$$



Appendix B

Report output

Output from ORA report generator

Node	Betweenness	Eigenvector	Degree
node1	0.0005	0.0035	0.0202
node2	0.0684	0.0093	0.0404
node3	0.0071	0.0027	0.0202
node4	0.0068	0.0029	0.0202
node5	0.0027	0.0023	0.0303
node6	0.0015	0.0018	0.0202
node7	0.0529	0.0045	0.0404
node8	0.0028	0.0017	0.0202
node9	0.0130	0.0047	0.0303
node10	0.0299	0.0131	0.0303
node11	0.0572	0.0199	0.0404
node12	0.0634	0.0131	0.0303
node13	0.0053	0.0015	0.0202
node14	0.0064	0.0015	0.0303
node15	0.0064	0.0012	0.0202
node16	0.0220	0.0032	0.0202
node17	0.0694	0.0118	0.0303
node18	0.0129	0.0037	0.0202
node19	0.0170	0.0032	0.0303
node20	0.0206	0.0078	0.0202
node21	0.2077	0.0283	0.0707
node22	0.1115	0.0256	0.0404
node23	0.0073	0.0089	0.0202
node24	0.0191	0.0101	0.0303
node25	0.0076	0.0036	0.0202

Node	Betweenness	Eigenvector	Degree
node26	0.0169	0.0045	0.0303
node27	0.0228	0.0088	0.0303
node28	0.0107	0.0055	0.0303
node29	0.0172	0.0057	0.0202
node30	0.0544	0.0184	0.0303
node31	0.0119	0.0057	0.0202
node32	0.0267	0.0048	0.0303
node33	0.0315	0.0074	0.0303
node34	0.0041	0.0058	0.0202
node35	0.0196	0.0177	0.0202
node36	0.0249	0.0109	0.0202
node37	0.0091	0.0032	0.0202
node38	0.0040	0.0021	0.0202
node39	0.0570	0.0197	0.0303
node40	0.0149	0.0062	0.0202
node41	0.0369	0.0054	0.0404
node42	0.0264	0.0102	0.0202
node43	0.0082	0.0191	0.0202
node44	0.4995*	0.0656*	0.0909
node45	0.3189	0.0405	0.0606
node46	0.0174	0.0111	0.0202
node47	0.0060	0.0040	0.0202
node48	0.0084	0.0052	0.0202
node49	0.0222	0.0168	0.0202
node50	0.4495	0.0625	0.1010*
node51	0.0205	0.0272	0.0303
node52	0.0002	0.0115	0.0202
node53	0.0524	0.0356	0.0404
node54	0.0311	0.0032	0.0303
node55	0.0501	0.0064	0.0303
node56	0.1185	0.0207	0.0404
node57	0.0086	0.0019	0.0202
node58	0.0071	0.0014	0.0303
node59	0.0124	0.0025	0.0202
node60	0.0380	0.0085	0.0303
node61	0.0061	0.0011	0.0202
node62	0.0105	0.0011	0.0303
node63	0.0191	0.0020	0.0202
node64	0.0354	0.0068	0.0202
node65	0.1562	0.0253	0.0505

Node	Betweenness	Eigenvector	Degree
node66	0.0273	0.0109	0.0303
node67	0.0224	0.0111	0.0303
node68	0.0480	0.0078	0.0303
node69	0.0088	0.0034	0.0202
node70	0.0107	0.0024	0.0202
node71	0.0032	0.0021	0.0202
node72	0.0348	0.0060	0.0455
node73	0.0257	0.0075	0.0354
node74	0.0345	0.0092	0.0505
node75	0.0753	0.0123	0.0505
node76	0.1023	0.0202	0.0303
node77	0.0068	0.0065	0.0152
node78	0.0069	0.0044	0.0202
node79	0.0184	0.0086	0.0303
node80	0.0019	0.0067	0.0202
node81	0.0634	0.0185	0.0303
node82	0.0282	0.0051	0.0202
node83	0.0113	0.0022	0.0202
node84	0.0029	0.0036	0.0152
node85	0.0000	0.0053	0.0202
node86	0.0097	0.0038	0.0202
node87	0.0015	0.0028	0.0202
node88	0.0298	0.0234	0.0303
node89	0.0266	0.0230	0.0303
node90	0.0088	0.0066	0.0202
node91	0.0052	0.0037	0.0303
node92	0.0143	0.0061	0.0202
node93	0.0013	0.0022	0.0202
node94	0.0172	0.0050	0.0202
node95	0.0607	0.0181	0.0303
node96	0.0109	0.0053	0.0202
node97	0.0054	0.0031	0.0202
node98	0.0381	0.0073	0.0404
node99	0.0130	0.0062	0.0202
node100	0.0258	0.0178	0.0202

Table B.1: Betweenness, Eigenvector, and Degree output from ORA risk report

Appendix C

Bar chart graphs

Bar chart quantities representation of nodes betweenness, eigenvector, and degree from large network (100 nodes).

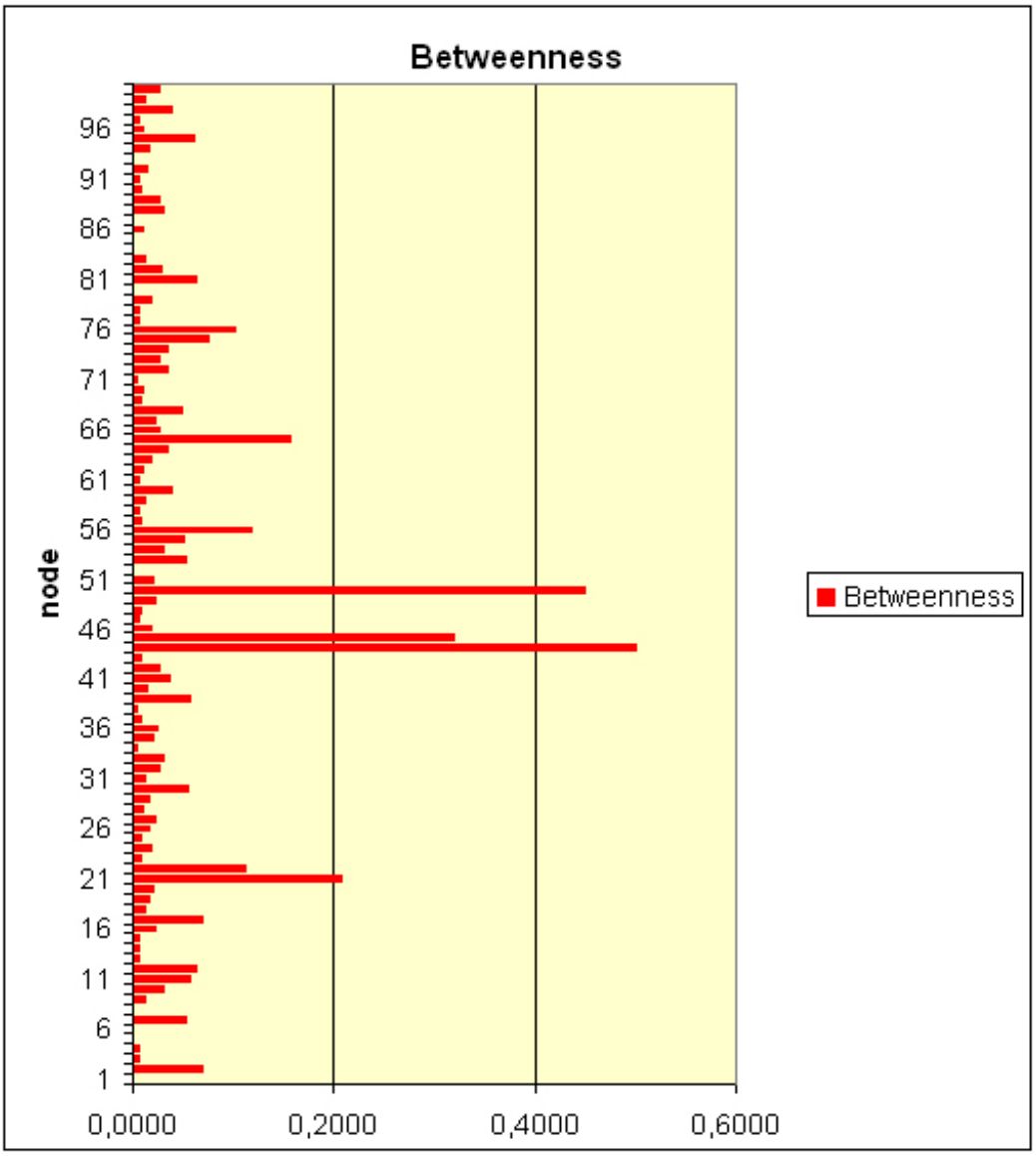


Figure C.1: Bar chart quantities representation of nodes betweenness from large graph

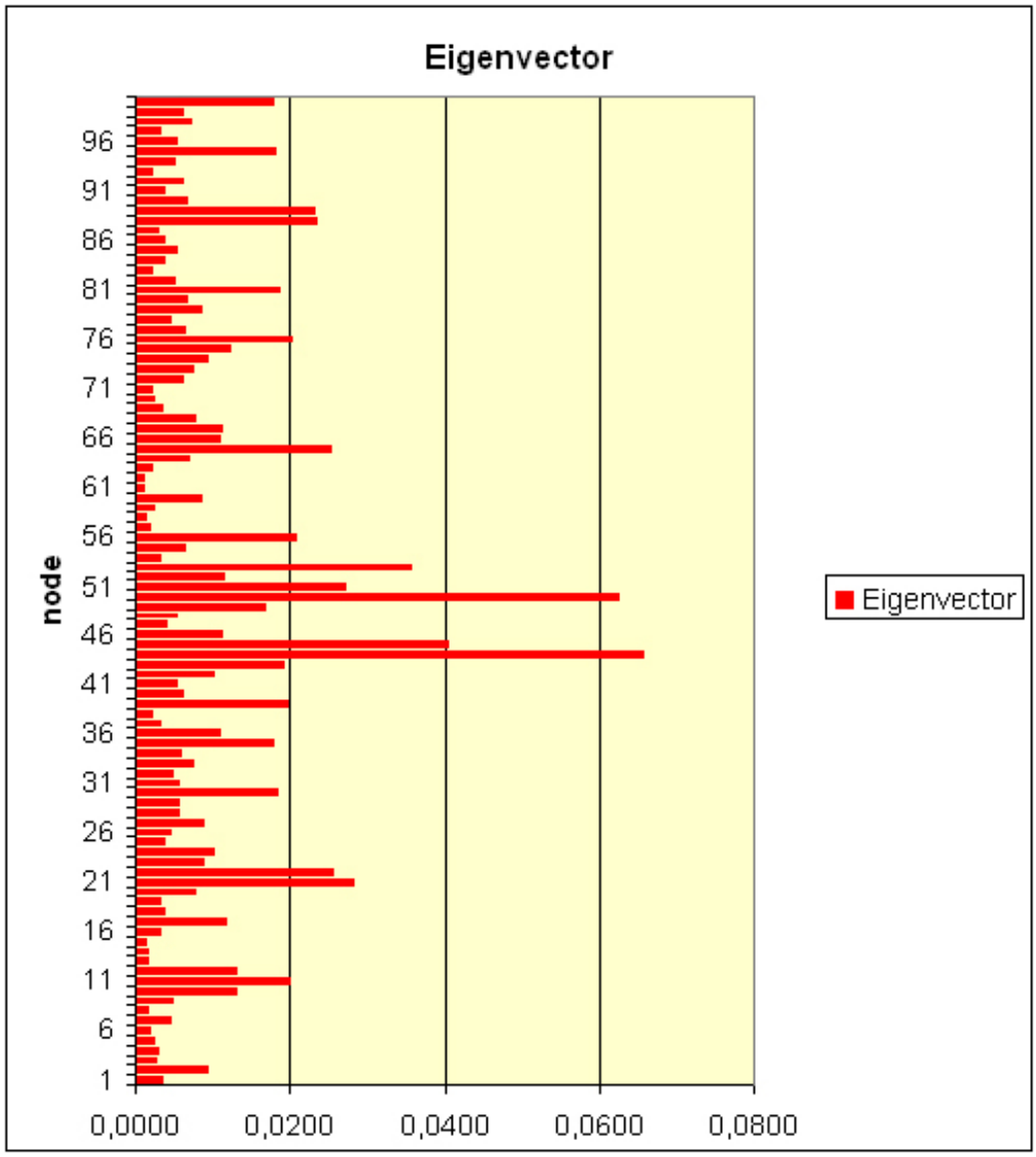


Figure C.2: Bar chart quantities representation of nodes eigenvector from large graph

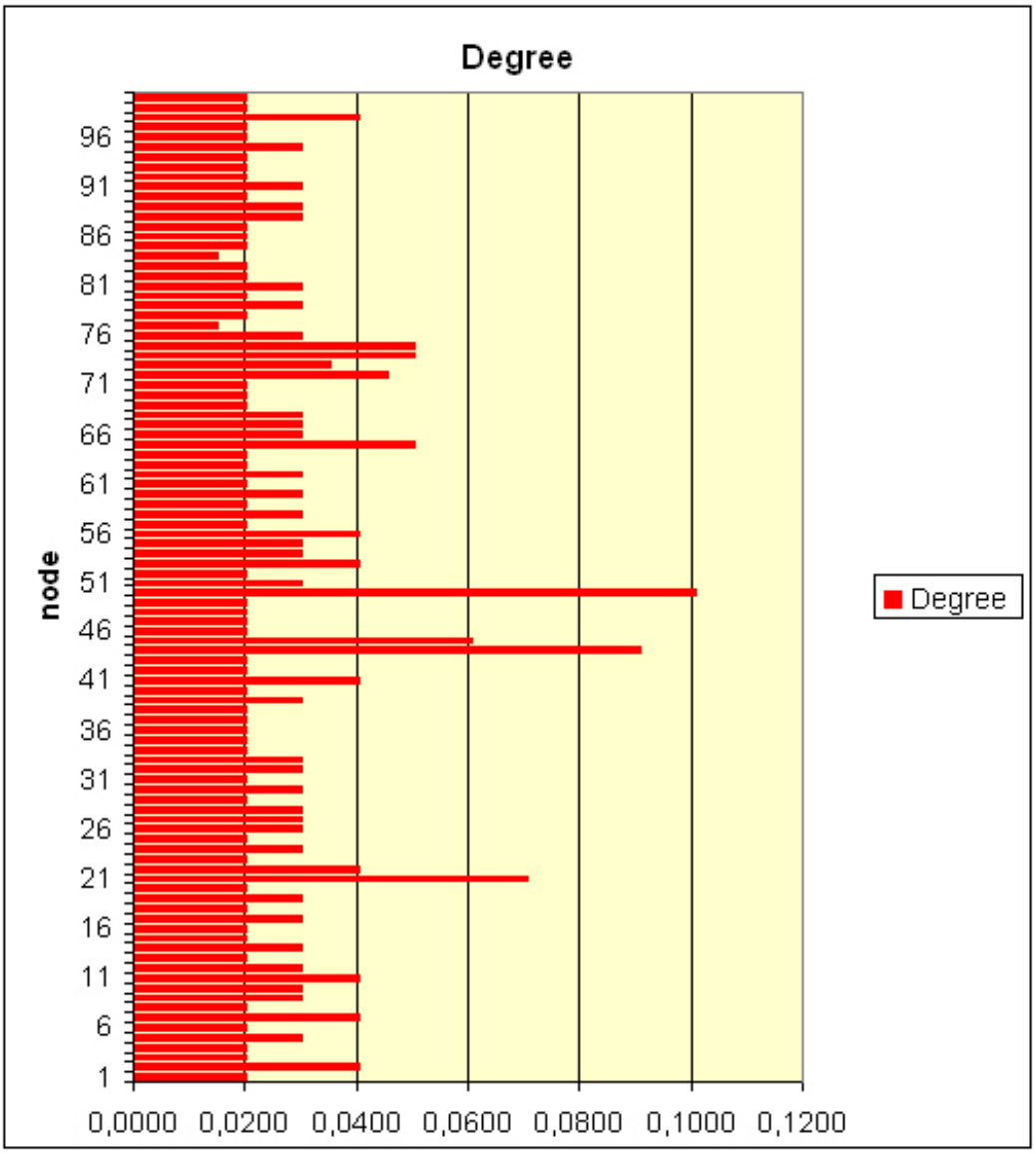


Figure C.3: Bar chart quantities representation of nodes degree from large graph

Appendix D

Tracing infections' movements

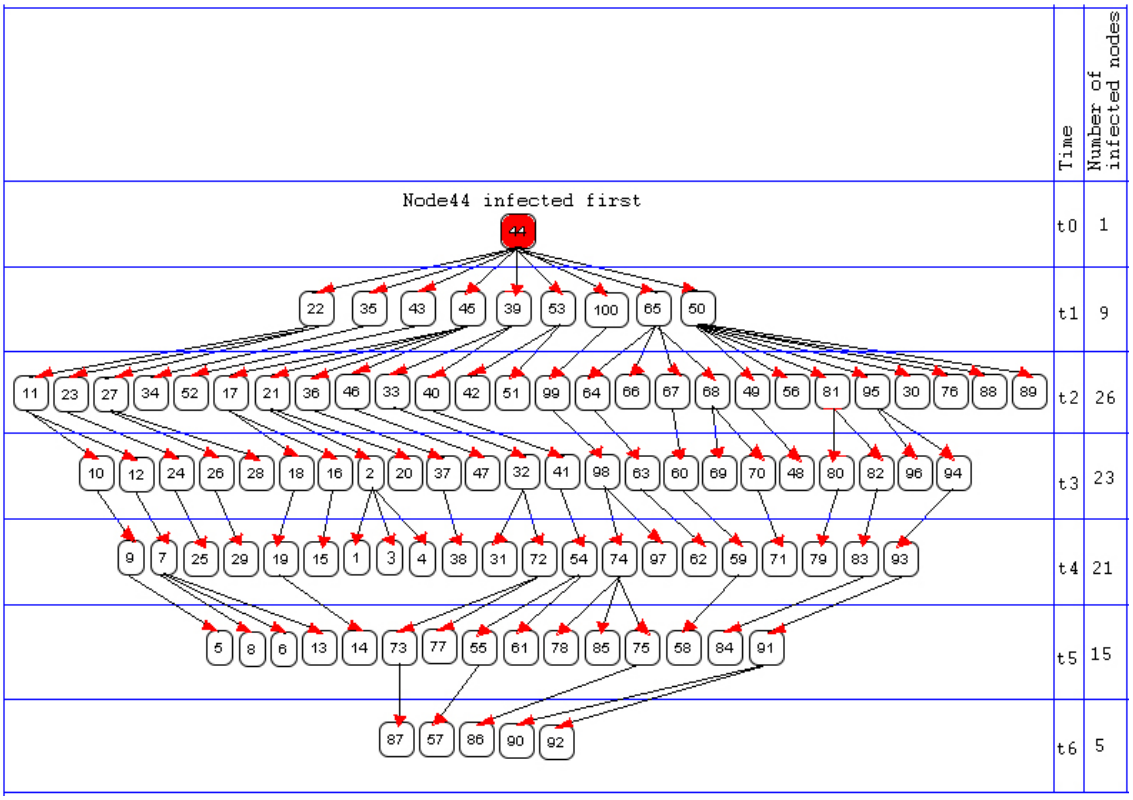


Figure D.1: Tracing infections' movements from $node44$

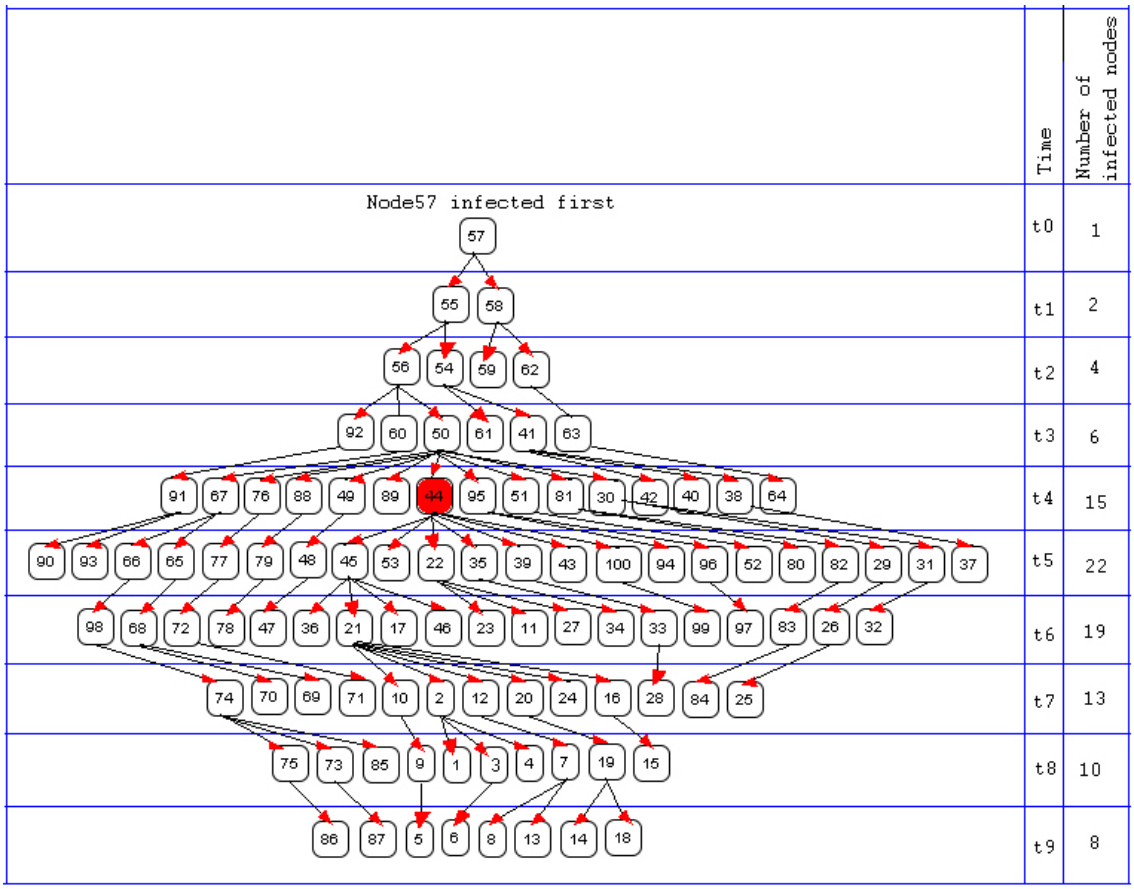


Figure D.2: Tracing infections' movements from $node57$

Appendix E

Eigenvector Ranking

We are ranking all nodes from most important nodes to less important nodes as follow:
{

1	Node44	0,0656
2	Node50	0,0625
3	Node45	0,0405
4	Node53	0,0356
5	Node21	0,0283
6	Node51	0,0272
7	Node22	0,0256
8	Node65	0,0253
9	Node88	0,0234
10	Node89	0,0230
11	Node56	0,0207
12	Node76	0,0202
13	Node11	0,0199
14	Node39	0,0197
15	Node43	0,0191
16	Node81	0,0185
17	Node30	0,0184
18	Node95	0,0181
19	Node100	0,0178
20	Node35	0,0177
21	Node49	0,0168
22	Node12	0,0131
23	Node10	0,0131
24	Node75	0,0123

25	Node17	0,0118
26	Node52	0,0111
28	Node46	0,0111
29	Node66	0,0109
30	Node36	0,0109
32	Node24	0,0101
33	Node2	0,0093
34	Node74	0,0092
35	Node23	0,0089
36	Node27	0,0088
37	Node79	0,0086
38	Node60	0,0085
39	Node68	0,0078
40	Node20	0,0078
41	Node73	0,0075
42	Node33	0,0074
43	Node98	0,0073
44	Node64	0,0068
45	Node80	0,0067
46	Node90	0,0066
47	Node77	0,0065
48	Node55	0,0064
49	Node99	0,0062
50	Node40	0,0062
51	Node92	0,0061
52	Node72	0,0060
53	Node34	0,0058
54	Node31	0,0057
56	Node28	0,0055
57	Node41	0,0054
58	Node96	0,0053
59	Node85	0,0053
60	Node48	0,0052
61	Node82	0,0051
62	Node94	0,0050
63	Node32	0,0048
64	Node9	0,0047
65	Node26	0,0045
66	Node7	0,0045
67	Node78	0,0044
68	Node47	0,0040
69	Node86	0,0038

70	Node91	0,0037
71	Node18	0,0037
72	Node84	0,0036
73	Node25	0,0036
75	Node69	0,0034
76	Node54	0,0032
77	Node37	0,0032
78	Node16	0,0032
80	Node97	0,0031
81	Node4	0,0029
82	Node87	0,0028
83	Node3	0,0027
86	Node5	0,0023
87	Node94	0,0022
88	Node83	0,0022
89	Node71	0,0021
90	Node38	0,0021
91	Node63	0,0020
92	Node57	0,0019
93	Node6	0,0018
94	Node8	0,0017
95	Node14	0,0015
96	Node13	0,0015
97	Node58	0,0014
99	Node62	0,0011
100	Node61	0,0011

}

Bibliography

[1] Susan Scott and Christopher J. Duncan. Biology of plagues: Evidence from historical populations. 2001.

[2] Thomas M. Chen and Jean-Marc Robert. Worm epidemics in high-speed networks. 2004.

[3] David Moore, Colleen Shannon, Geoffrey M. Voelker, and Stefan Savage. Intrnet quarantine: Requirements for containing self-propagating code. 2003.

[4] Zoran Nikoloski, Narsingh Deo, and Ludek Kucera. Correlation model of worm propagation on scale-free graphs. October 01, 2005.

[5] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the inter topology. 1999.

[6] Mark Burgess. Analytical network and system administratin: Managing human-computer system. pages 73 – 84, 2004.

[7] Glyn James, David Burley, Dick Clements, Phil Dyke, John Searl, and Jerry Wright. Modern engineering mathematics. pages 337 – 346, 2001.

[8] M. E. J. Newman. Exact solution of epidemic model on networks. November 28, 2001.

[9] Helen Trottier and Pierre Philippe. Deterministic modeling of infections diseases: Theory and methods. 1(2), 2001.

[10] Zesheng Chen, Lixin Gao, and Kevin Kwiat. Modeling the spread of active wormst. 2003.

[11] Denis Mollison. Epidemic models: Their structure and relation to data. page 18, 2003.

[12] Tim Daniel and Peter Bleckmann. Epidemic algorithms. August 4th, 2004.

[13] Hiroyuki Okamura, Hisashi Kobayashi, and Tadashi Dohi. Markovian modeling and analysis of internet worm propagation. 2005.

[14] Hethcote Herbert. The mathematics of infectious diseases. pages 599 – 653, 2000.

[15] R. Pastor-Satorras and A. Vespignani. Epidemics and immunization in scale-free networks. May 14, 2002.

[16] Shih Ching Fu. Realism in epidemic models. 2002.

[17] Alan Solomon. Epidemiology and computer viruses. 1990.

[18] Stuart Staniford, Vern Paxsony, and Nicholas Weaver. How to own the internet in your spare time.

[19] Christopher Griffin and Richard Brooks. A note on the spread of worms in scale-free networks. 2006.

[20] Yang Wang and Chenxi Wang. Modeling the effects of timing parameters on viruse propagation.

[21] Matthew M. Williamson. Throttling viruses: Restricting propagation to defeat malicious mobile code1. December 10th, 2002.

[22] M. E.J. Newman. A measure of betweenness centrality based on random walks.

[23] Goh, E K I, B. Kahng OH, and D. Kim. Betweenness centrality correlation in social networks. 2003.

[24] Geoffrey Canright and Keneth Engoe-Monsen. Spreading on networks: a topographic view. 2001.

[25] Jari Saram and Kimmo Kaski. Modeling development of epidemics with dynamic small-world network. 2000-2005.

[26] Mathematica at: http://www.wolfram.com/products/mathematica/introduction.html.

[27] Excel at: http://office.microsoft.com/en au/FX010858001033.aspx.

[28] Kathleen M. Carley and Jeff Reminga. Ora: Organization risk analyzer: Casos technical report. 2004.

[29] Marian Boguna, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic spreading in complex networks with degree correlations. 2003.

[30] Yang Wang, Deepayan Chakrabarti, Chenxi Wang, and Christos Faloutso. Epidemic spreading in real networks: An eigenvalue viewpoint. 2003.

[31] Romulus Breban, Raffaele Vardavas, and Sally Blower. Linking population-level models with growing networks: A class of epidemic models. 2005.

[32] Li-Chiou Chen and Kathleen M. Carley. The impact of countermeasure propagation on the prevalence of computer viruses. 2004.

([1])

[1] Author of this paper (Epidemic Network and Centrality): Akram Rustam: akram.hr@hotmail.com