Internet Topology Characterization on AS Level

SARA ANISSEH

KTH Electrical Engineering

Master's Degree Project
Stockholm, Sweden

XR-EE-LCN 2012:013

KTH Electrical Engineering

THE INTERNET TOPOLOGY CHARACTERIZATION ON AS LEVEL

SARA ANISSEH

INTERNAL SUPERVISOR:

ASSOCIATE PROFESSOR VIKTORIA FODOR
School of Electrical Engineering in Kungliga Tekniska Högskolan (KTH)

EXTERNAL SUPERVISOR:

PROFESSOR LJILJANA TRAJKOVIC
School of Engineering Science in Simon Fraser University (SFU)

SIMON FRASER UNIVERSITY
THINKING OF THE WORLD

Abstract

This study investigates the Internet topology characterization on AS level driven from Border Gateway Protocol (BGP) tables which are collected from Réseaux IP Européens (RIPE) datasets during the seven year period, from 2003.07.30 to 2010.07.30.

The investigation shows that despite of the growth of the Internet with lack of centralized control, some properties of the Internet follow certain rules and some properties remain the same during years. It demonstrates that the Internet, on AS level, exists in the form of clusters of ASs and the connected ASs with higher connectivity become even more connected during time. The spectral analysis of adjacency and normalized Laplacian matrix shows that the eigenvalues of both matrixes follow power-laws with high correlation coefficient with no considerable change in exponent values during years.

Acknowledgment

I would like to thank my supervisors: Professor Ljiljana Trajkovic and Assosiate Professor Viktoria Fodor for their support, encouragement and patience.

Table of Figures

Contents

1. Introduction

1.1 Overview

The Internet, the network of networks [1] with millions of users, which are connected to each other by different technologies, is growing every day. It is of great interest to scientists to analyze certain topological behaviors and characteristics [2] [3][4][5][6] of the Internet infrastructure despite of all the various applied technologies in conjunction with its rapid growth without a central control [6]. Determining and mapping certain characteristics of the Internet topology helps us to have a better understanding of the Internet infrastructure therefore it makes it possible to model or simulate the Internet for further studies and consequently to better solutions for problems and new protocols [2]. Due to the large size of the Internet, it is preferred to investigate the research on the representatives of groups of routers which are called Autonomous Systems (AS). In order to analyze the Internet topology on AS level it is needed to extract the Internet data at AS level with the help of available datasets on the Internet. In this study analyzing the Internet data at AS-level will be achieved by generating a graph corresponding to connected ASs collected from Border Gateway Protocol (BGP) tables since it is possible to extrapolate the topological behavior of a graph from its related matrixes. This study shows that despite many factors, dynamics and of all the differences in applied technologies of networks, including all changes and rapid growth in the Internet some characteristics of the Internet remain the same and that in fact they follow specific rules over time. Literature review reveals that the existence of several power-laws in the Internet graph has been proved. For instance the presence of power-laws in the following properties has been seen:

- Node degree vs. node rank
- Node degree frequency vs. degree
- Number of nodes within a number of hops vs. number of hops
- eigenvalues of the adjacency and normalized Laplacian matrix vs. the order of the eigenvalues [3][4] [6][8]

On the other hand, the spectrum of adjacency and normalized Laplacian matrix reveal more topological characteristics of the Internet such as the diameter of the network, presence of cohesive clusters, long paths and bottlenecks, connectivity, and the randomness of a graph [2]. It is possible to show the clustering properties of ASs by calculating the eigenvector corresponding to the largest eigenvalue of Laplacian matrix [4]. The second smallest eigenvalue of Laplacian matrix or algebraic connectivity specifies the connectivity characteristic of a graph. It is shown that the Internet exists as clusters of connected ASs and over time, the more connected ASs become even more connected.

Therefore, regardless of how rapidly and uncentralized the Internet grows, the Internet infrastructure still follows certain rules and several power-laws have been found in the Internet graph topology.

1.2 Goals

This thesis collected information of BGP (Border Gateway Protocol) routing tables on each AS from the dataset in RIPE (Réseaux IP Européens, which is French for European IP Networks) project over a period of seven years, from 2003.07.30 to 2010.07.30. The Internet graph at AS level has been mapped with the data taken from information in BGP routing tables. The adjacency and normalized Laplacian matrix of related graphs has been made and consequently the spectrum of a graph based on adjacency and normalized Laplacin matrix has been calculated.

In this thesis we prove that eigenvalues of adjacency matrix and normalized Laplacian matrix follows power-law without considerable change in power-law exponents during a period of seven years up to 2010.07.30, this occurs despite of many changes to the Internet in terms of its growth.

This thesis also shows that the Internet graph exists as clusters of connected ASs and over increasing time the ASs which are already more connected than other ASs actually become more connected.

For the purpose of this thesis all the calculations and depicted graphs have been done in MATLAB

1.3 Outline

Chapter 2, 3, 4, and 5 explains the required background for this report. In chapter 2 different routing protocols are explained. Chapter 3 specifically discuss about the Internet routing protocol and introduction to the datasets which is used in this study. Chapter 4 is some basic concepts about graph theory and the definition of spectral analysis. Chapter 5 explains the required knowledge of power-laws for this study. Chapter 6 and 7 explains the practical part of the study. Chapter 6 specifically explains how to collect the real information of BGP tables from BGP routers and how to convert it to human readable format. Chapter 7 demonstrates the analysis of collected information from BGP tables and finally chapter 8 is the conclusion of the study.

2. Basic Internet Structure

Due to the size and scope of the Internet it is not possible for one routing protocol to take care of all the updates in routing tables. Therefore, “the Internet is divided into autonomous systems (ASs) [7].” “The Autonomous system (AS) is a group of networks and routers under the authority of a single administration [7].” The routing protocols are divided into two major groups: Intradomain Routing Protocol which is the routing inside the autonomous system and Interdommain Routing Protocol which is the routing between the autonomous systems.

2.1. Internal Routing Protocol (Intradomain Routing Protocol)

In this subchapter we focus on two important internal routing protocols: Distance Vector Routing Protocol and Link State Routing Protocol.

2.1.1. Distance Vector Routing Protocol

Distance vector routing protocol applies the Bellman-Ford algorithm for implementation. In this case routers do not have information about the whole path to the destination. “The route with minimum distance can be considered the least cost route between any two nodes [7].” In each router there is a table, wherein lies a list of nodes and the minimum cost to reach them. This information will be shared to the neighbors.

In initiation state the routers only have the information cost of their immediate neighbors but after a while they will receive more information of farther nodes from the immediate neighbors which are acting as intermediate nodes. This sharing of information between the neighbors is done whenever there is a change or update in the table, or it can also be done periodically, alternatively it can also be called Triggered update and Periodic update. One of the famous examples of distance vector routing protocol is Routing Information Protocol (RIP). The metric in RIP is on a hop count basis which means that the number of links to final destination is the counted distance. RIP has three timers: Periodical Timer, Expiration Timer and Garbage Collection Timer.

Periodical Timer

“Every 30 seconds a RIP router will broadcast a lists of networks and the subnets it can reach periodically [7]” regardless of whether or not the information has changed.

Expiration Timer

A route can be received to a router from the update information. “After receiving the updates regarding the route a time is set to 180 which is an expiration timer [9].” For each single received update the expiration timer will be set. If a problem occurs and no update is received during expiration time the route is considered expired and the hop count is set to 16 which means the destination is unreachable.

Garbage Collection Timer

If the information of a router is invalid then it will not erase this information immediately from the routing table, rather it will keep it in its table for 120 seconds with metric of 16 and after this 120 second period it will erase the information. This 120 second is referred to as the garbage collection timer.

A router only knows information about its own routing table and its neighbors routing table. Every 30 seconds (periodical update) or for each update (triggered update) the router will reevaluate its table to check if the update route has less cost, and if this condition is true then it will update the routing table otherwise it will not make any changes. Accordingly the routing table will be set for all the routers over the time to interconnect and process any changes in the routing tables.

2.1.2. Link State Routing Protocol is based on Dijkstra algorithm. “In this Protocol each node floods the information to all other nodes in the network. Contrary to distance vector routing protocol, each node in the domain has the entire topology of the domain like the list of nodes, links, how they are connected including the type, cost and the condition of the links (up or down) [7].”

The Open Shortest Path First (OSPF) protocol is an intradomain routing protocol based on link state routing and its domain is also an AS. The OSPF protocol allows the administrator to assign a cost, called the metric, to each route. The metric can be based on a type of service (minimum delay, maximum throughput, and so on). As a matter of fact, “a router can have multiple routing tables: each based on a different type of service [7].”

The OSPF Protocol does not impose a hop-count restriction so it is suitable for larger networks.

The protocol uses small "hello" packets to verify link operation without transferring large tables every 10 seconds and if after 40 seconds no “hello” packet is received from a neighbor router it is considered as a dead interval. “OSPF uses the “hello” message to create neighborhood relationships and to test the reachability of neighbors [7].” Before a router routes a packet it processes the information about its neighbors to check if they are reachable or alive. This can be done by sending the "hello" packet to the neighbor. OSPF uses the shortest path algorithm which uses a tree that contains network topology.

In the beginning the router sends the "hello" packet to make the routing connection. "Link state" is another message used for updating in specified intervals so that all the routers know the routing information of the entire network. Then the shortest path can be extracted from "shortest path tree".

2.2. External Routing Protocols (Interdomain Routing Protocol)

It is not feasible to use intradomain routing protocols between the ASs. Distance vector routing will be instable for a large number of hops. On the other hand link state routing needs a large amount of resources to calculate routing tables [1] and also traffic in the network will increase dramatically because of flooding. In order to deal with the problem, another protocol is used for interdomain routing. Path Vector Routing Protocol is a suitable protocol for external routing. The idea is similar to distance vector routing but with some differences. In path vector routing there is a concept of speaker node which is a node in autonomous system that acts on behalf of the entire autonomous system. In this case speaker nodes talk to each other and create the tables and instead of advertising the metric, they advertise the path. Initially, each speaker node has the information of its own autonomous system's nodes but then they start sharing their own tables with their neighbors and over time they will update their tables with the possible paths to other ASs. There are several metrics to find the optimal path to the destination. It's not just the number of intermediate ASs but also factors of security, reliability and other metrics are important in choosing the optimal path.

To be able to send a datagram from a source to destination, two different kinds of routing tables are defined. One table for the intradomain routing inside an AS and the other table for interdomain routing between the ASs. A datagram can find the route to destination by the help of interdomain tables from one AS to another AS. When it reaches the destination AS, it does not need to travel from one AS to another AS, instead, it needs to find the destination inside the AS. As it was mentioned before intradomain routing protocols are used inside an AS. Therefore, after reaching the destination AS, the nodes use the intradomain routing tables in order to find the destination.

3. BGP

Border Gateway Protocol is one of the famous interdomain routing protocols in TCP/IP networks which uses path vector routing. It also uses CIDR (Classless Interadomain Addressing) notation for addressing in order to reduce the size of routing tables by grouping the routes together. “The exchange of routing information between two routers using BGP takes place in a session [7].” “A session is a connection that is established between two BGP routers only for the sake of exchanging routing information [7].” BGP sessions are TCP based and they last for a longer time until something unusual happens therefore they can also be called as semi-permanent connections. There are two types of BGP sessions: internal (I-BGP) and external (E-BGP). E-BGP is used between the two speaker nodes of different ASs since I-BGP is used between the routers inside an autonomous system. BGP routers exchange routing information using four types of messages [9]: open, update, keepalive, notification.
In order to make a TCP connection with a neighbor, the router sends an open message to its neighbor and if the neighbor accepts the connection, it will send a keepalive message to the router and then the connection between two routers are established. In order to transfer the updates to the neighbors update message will be used. When a router wants to close a connection or whenever an error occurs, a notification message will be sent.

3.1. Autonomous System

“The Internet is divided into hierarchical domains called autonomous systems [7].” “A large corporation that manages its own network and has full control over it is an autonomous system [7].” As an example a local ISP (Internet Service Provider) can be an AS. There are three types of Autonomous Systems:

- **Stub AS**

Stub AS has only one connection to another AS. It means that the traffic will not pass through the Stub AS. It can either receive the traffic or it can send the traffic to one AS. “A stub AS is either a source or a sink [7].”

- **Multihomed AS**

Multi-homed AS has more connection to other ASs but still it does not transit the traffic through itself to another AS. It can also send the traffic to more than one AS and also receive traffic from more than one AS but still it is a source or a sink.

- **Transit AS**

Transit AS has multiple connections to multiple ASs and it can also transit the traffic through itself. As an example national and international ISPs (Internet Service Provider) can be a transit AS.

Each AS has a unique number which represents its network uniquely on the Internet. This number is called ASN (Autonomous System Number). IANA (Internet Assigned Numbers Authority) is the organization responsible for assigning the AS numbers. Initially, IANA defined AS numbers by 16 bits but as the Internet is growing rapidly, IANA started to use 32 bits for ASNs on December 1st 2006. By using 32 bits they expanded the pool size from 65536 to 4294967296.

In 16-bit format each ASN has been represented by 16 bits from a pool of 65536. Out of this pool, 1023 numbers are reserved for private use, 3074 numbers are reserved for special use and the remaining of 61439 numbers is available for use to support the Internet's public inter-domain routing system [10].

In 32-bit format there is a pool of 4294967296. From this pool 1023 numbers (which are common in 16-bit as well) are reserved for private use, 65537 are reserved for special use (just in 32-bit, therefore in total there are 65537 + 3074 = 68611 numbers of reserved for special used in 32-format) and the remaining pool of 4294897662 numbers are available for use to support the Internet's public inter-domain routing system [10].

The available numbers can be either allocated or unallocated. Not all the available numbers are allocated but the allocated ones are managed by Regional Internet Registries (RIRs). There are five RIRs around the world:

- AfriNIC (African Network Information Center) is the RIR for Africa.
- APNIC (Asia Pacific Network Information Centre) is the RIR for the Asia Pacific region.
- ARIN (American Registry for Internet Numbers) is the RIR for Canada, many Caribbean and North Atlantic islands, and the United States.
- LACNIC (Latin America and Caribbean Network Information Centre) is the RIR for the Latin American and Caribbean regions.
- RIPE-NCC (The Réseaux IP Européens Network Coordination Centre) is the RIR for Europe, the Middle East and parts of Central Asia.

3.2. BGP Dataset

BGP datasets collect and store the information of routing tables in MRT format. In this thesis project datasets form RIPE (Réseaux IP Européens) project is used. Some terms is explained below in order to make a better understanding of the BGP datasets which are used in this study and then in chapter 6 it will be discussed in more details.

RIR: Regional Internet Registries are the organizations which manage the allocated ASs in their own region. As it was mentioned in 3.1 there are five different RIRs for five different regions which are AfriNIC, APNIC, ARIN, LACNIC, and RIPE-NCC.

RIPE NCC: The Réseaux IP Européens Network Coordination Centre is the Regional Internet Registry (RIR) for Europe, the Middle East and parts of Central Asia. RIPE NCC manages the allocated ASs in the mentioned regions. The head quarter is in Amsterdam, Netherlands [11]. The start point of RIPE NCC was in April 1992.

RIS: Routing Information Service is a project in RIPE NCC that "collects and stores Internet routing data from several locations around the globe. RIS offers tools that bring this data to the Internet community [12]."

RRC: RIPE NCC uses Remote Route Collectors (RRC) in different part of the world to collect the Internet routing data. "It is a software router, running on a Linux platform that only collects default free BGP routing information [13]." The RRCs collect the routing data from different regions. There are 17 RRCs in different geographical place to collect the BGP tables in specific times. As it is shown below, most of them are located in Europe:

- rrc00.ripe.net at RIPE NCC, Amsterdam,
- rrc01.ripe.net at LINX, London
- rrc02.ripe.net at SFINX, Paris.
- rrc03.ripe.net at AMS-IX, Amsterdam.
- rrc04.ripe.net at CIXP, Geneva.
- rrc05.ripe.net at VIX, Vienna.
- rrc06.ripe.net at Otemachi, Japan.
- rrc07.ripe.net in Stockholm, Sweden.
- rrc08.ripe.net at San Jose (CA), USA.
- rrc09.ripe.net at Zurich, Switzerland.
- rrc10.ripe.net at Milan, Italy
- rrc11.ripe.net at New York (NY), USA.
- rrc12.ripe.net at Frankfurt, Germany.
- rrc13.ripe.net at Moscow, Russia.
- rrc14.ripe.net at Palo Alto, USA.
- rrc15.ripe.net at Sao Paulo, Brazil.
- rrc16.ripe.net at Miami, USA.

MRT Format: The Internet routing data in each RRC is stored according to the collected dates and times in MRT format. "The MRT format was developed to encapsulate, export, and archive the information regarding network behavior analysis (by studying routing protocol transactions and routing information base snapshots) in a standardized data representation [14]." "MRT routing information export format represents an effective way of storing BGP routing information in binary dump file [15]."

Route Views: Route Views is a project founded by Advanced Network Technology Center at the University of Oregon to allow. The Route Views BGP monitor collects update streams from 25 BGP speaking neighbors. It collects five to six million updates per day [16]. Like RIPE project, it also collects and restores BGP routing table information. Most information is collected from North America.

According to the explanation, RRCs which are located in 17 different part of the world are responsible to collect the BGP table information at specific times during a day and store the information in MRT format. The MRT files, which are the BGP routing table information, should be converted to human readable files in order to understand the information of the routing tables and investigate the properties of the extracted data. There are some tools to convert the MRT files to ASCI and human readable file. These tools are presented in both RIPE and Rout Views project. This project converted the MRT files to human readable files with the tools introduced in both RIPE and Rout Views project in order to compare the converted files from different introduced tools. The comparison between the converted files shows the same result. In this project with the help of the tools in both Ripe and Route views project we could convert the MRT files to human readable files.

Route Views project introduces three tools to convert the binary MRT files to ASCI which are:

- zebra-dump-parser
- bgpdump
- route_btoa

This study used zebra-dump-parser from Route Views project and pgbdump from RIPE project to convert the MRT files to ASCI.

zebra-dump-parser: is a perl script written by Marco d'Itri. This script is capable of parsing the MRT files.

bgpdump: “bgpdump is in the libbgpdump sources which are maintained by RIPE RIS [17].” *Libbgpdump*: Libbgpdump is a C library designed to help analyzing dump files which are produced by Zebra/Quagga or MRT.

Below is an example of MRT file syntax after conversion to human readable format:

TIME: 07/31/04 00:00:00

TYPE: TABLE_DUMP/INET

VIEW: 0

SEQUENCE: 0

PREFIX: 3.0.0.0/8

FROM:195.66.224.99 AS13237

ORIGINATED: 07/30/04 10:19:38

ORIGIN: IGP

ASPATH: 13237 1299 1239 80

NEXT_HOP: 195.66.224.99

COMMUNITY: 13237:44049 13237:46021

STATUS: 0x1

4. Internet Topology and Graph Theory

4.1. Basic Graph Theory

The Internet can be presented as a huge undirected graph on AS level where routers are represented by nodes which are also called vertices and transmission lines represented by edges which are the connection between the nodes [2]. Therefore the Internet can be presented as G (V, E) where V represents the vertices or nodes and E represents edges or the connection between nodes. Like all undirected graphs, the Internet graph can also be presented by different kinds of matrixes. Among all representative matrixes, adjacency, diagonal, Laplacian and normalized Laplacian matrixes are of more importance.

The Internet graph can be represented by an adjacency matrix of size V×V. Two nodes are called adjacent if they are connected by an edge together. Therefore an adjacency matrix can be created based on adjacent node definition which means that adjacency matrix is a way to show which nodes are adjacent to each other. The network can be presented by an adjacency matrix as below:

$$A(i,j) := \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases}$$ **(Formula4. 1)**

In an undirected graph the number of edges from a certain node is called node degree. To present the node degree in matrix format a diagonal matrix should be defined. In order to make the diagonal matrix out of the adjacency matrix, we should sum up the values in each row of the adjacency matrix and set them as the diagonal values of the diagonal matrix, all the other elements except the values on diagonal are set to zero. This diagonal shows the degree of each corresponding node or router. The mathematical representation can be shown as below:

D(G)= diag(sum(A(G))) **(Formula4. 2)**

One of the most important matrixes in spectral graph theory is Laplacian matrix which is also called admittance matrix or Kirchhoff matrix. Outstanding properties of Laplacian matrix reveal the behavior of the graph. Some of the most important properties of graph are extrapolated by Laplacian matrix, for instance connectivity and clustering are defined by Laplacian matrix which will be explained further more. Laplacian Matrix can be represented as below, where D(G) is the corresponding diagonal matrix of the graph and A(G) represents the adjacency matrix:

L(G)= D(G) − A(G) **(Formula4. 3)**

According to definition above this matrix is closely related to adjacency matrix. Another representation of Laplacian matrix can be defined as below where $\deg(i)$ is the degree of node i:

$$L(i,j) := \begin{cases} \deg(i) & \text{if } i = j \\ -1 & \text{if } i \neq j \text{ and } v_i \text{ is adjacent to } v_j \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(Formula4. 4)}$$

Based on Laplacian matrix L and diagonal matrix D it is possible to define the normalized Laplacian matrix:

$NL(G) = D^{-1/2} L D^{-1/2}$ **(Formula4. 5)**

Another presentation of normalized Laplacian matrix can be defined as below.

$$NL(i,j) := \begin{cases} 1 & \text{if } i = j \text{ and } d_i \neq 0 \\ -\frac{1}{\sqrt{\deg(i)\deg(j)}} & \text{if } i \text{ and } j \text{ are adjacent} \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(Formula4. 6)}$$

All the eigenvalues of normalized Laplacian matrix are real and non-negative [29]. One of the useful properties of Normalized laplacian matrix is that the ranges of all the eigenvalues are between 0 and 2 so it makes comparing the spectra of two graphs easier [32]. The smallest eigenvalue of normalized Laplacian matrix is always 0. "Another motivation for using the normalized Laplacian Matrix is to make it more naturally to deal with non-regular graphs since it behaves more naturally with non-regular graphs [18]" which means that "in some situations the normalized Laplacian is a more natural tool that works better than the adjacency matrix or combinatorial Laplacian to analyze the properties of a graph [18]." "The eigenvalues of the normalized Laplacian matrix are in a "normalized" form [29][30]" and they relate well to other graph invariants for general graphs in a way that the other two matrixes (Adjacency and Laplacian) fail to do [30].It is consistent with the eigenvalues in spectral geometry and in stochastic processes which is an advantage of normalized Laplacian matrix [30].

4.2. Eigenvalues and eigenvectors:

Eigenvalues and eigenvectors are two important concepts in spectral analysis of a graph. In other words spectral analysis of a graph based on eigenvalues and eigenvectors. In order to understand eigenvalues and eigenvectors and their specific characteristic which makes them to

reveal the important properties of the graphs, we need to know some concepts like: square matrix, transpose of a matrix, symmetric matrix,

Square matrix: every n by n (n × n) dimension matrix is called a square matrix.

Transpose of a matrix: Consider we make a matrix A^T in which all the columns in matrix A are the corresponding rows in matrix A^T and all the rows in matrix A are the corresponding columns in matrix A^T. Matrix A^T is called a transpose matrix of matrix A.

Symmetric matrix: If a square matrix equals to its transpose matrix we call it a symmetric matrix. In other words A is symmetric if $A= A^T$.

Eigenvectors are also known as proper vectors, characteristic vectors or latent vectors which are set of vectors associated with matrixes [19].The term eigenvector is meaningless without eigenvalue. The both definitions rely on each other. For each set of eigenvector there is a corresponding eigenvalue. "Eigen means self in German [20]" and the reason of choosing this name relies on eigenvalue and eigenvector concept. If a non-zero vector v multiplied by matrix A and the result of the multiplication still remains parallel to the original vector, or in other words the result of the multiplication of A by v equals to λv, where λ is an scalar, we call the v vector as the eigenvector of matrix A and λ as the eigenvalue of the corresponding eigenvector v of the matrix A. It can be represented as below

$$Av = \lambda v \qquad \textbf{(Formula4. 7)}$$

Based on the definition of eigenvalues, the spectrum of a matrix is defined. The whole set of eigenvalues of a matrix is called the spectrum of a matrix. Therefore spectral analysis of a matrix relates to investigation of eigenvalues of a matrix since they reveal some important characteristics of a matrix. Some of these characteristics are listed below:

1- Number of connected components in a graph which is defined as the number of times that eigenvalue of a Laplacian matrix becomes zero. It shows the number of disconnected sub-networks [21] [22]. In other words the multiplicity of 0 as an eigenvalue of L is the number of connected components of the graph [30].

2- Spectral gap: The smallest non-zero eigenvalue of the Laplacian matrix is called spectral gap which is the difference between two largest eigenvalues.

3- The second smallest nonzero eigenvalue in Laplacian matrix is called algebraic connectivity or Fiedler value [6][23]. If and only if the graph is connected, the algebraic connectivity (second smallest eigenvalue of Laplacian matrix) will be greater than zero. The magnitude of the algebraic connectivity shows the robustness of the graph or network. The bigger algebraic connectivity values show that the more connected is the graph or network and it is more difficult to make the graph into disconnected components. The algebraic connectivity depends on both the number of vertices and edges.

4-The eigenvector corresponding to the largest eigenvalues shows the clustering of a graph or network [2][3][4][5][6]. In general the eigenvectors corresponding to the largest eigenvalues reveal the global attributes like clustering of nodes in a graph or network and the eigenvector corresponding to the small eigenvalues reveal the local attributes like connectivity. “Clustering means partitioning of a graph, so that the edges between different groups have low similarity (low distance) and the edges within a group have high weight (low distance) [31].” In other words clustering is the task of grouping the vertices of the graph into clusters [32] in which there are many edges within each cluster and few edges between the clusters.

5- Symmetric normalized Laplacian matrixes are positive semi-definite and have n non-negative real-valued eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$.

6- The multiplicity k of the eigenvalue 0 of symmetric normalized Laplacian matrix equals the number of connected components $A, \ldots, A_k$ in the graph [31].

5. Power-laws and The Internet Structure

A power-law is a mathematical relationship between two quantities. "When the frequency of an event varies as a power of some attribute of that event, the frequency is said to follow a power law [24]." In other words, "a relationship in which a relative change in one quantity gives rise to a proportional relative change in the other quantity, independent of the initial size of those quantities [25]." "Power-laws are expressed in the form of $y \propto x^a$, where y and

x are the measures of interest and exponent a is a constant [6]." Therefore "the outstanding characteristic of power-law is its scale invariance which means that a feature of objects or laws that does not change if scales of variables, are multiplied by a common factor [26]."

Existence of power-laws can be shown by $y \propto x^a$ which means that y is proportional to x to

the power of a, therefore the presence of power-laws demonstrates a regularity. In order to provide a schematical view of power-laws, we need to depict the power-laws equation. Converting the power-laws relation ($y \propto x^a$) to an equation can be done by a multiplication

of a constant 10^b:

$y = 10^b \times x^a$ (where b is a constant, therefore 10^b will be a constant)

$\log y = \log 10^b + \log x^a$ **(Formula5. 1)**

$\log y = b + a \times \log x$

$y' = b + ax'$

According to above y and x has a linear relation in log-log scale. Normally, plotting the power-law relations in logarithmic axes gives us a linear relationship therefore any two quantities which have a linear relation with each other on logarithmic scale can be stated as power-law relation.

"Power-laws are very important because they reveal an underlying regularity in the properties of systems [25]." Often highly complex systems have properties where "the changes between phenomena at different scales are independent of which particular scales we are looking at [25]."

In this thesis project shows that power-law relation exists in eigenvalues of adjacency matrix and normalized Laplacian matrix versus node index. We are going to prove that this equation $\lambda_i \propto i^{\varepsilon}$ exists if i represents index and λ_i represents the adjacency or normalized

Laplacian matrix eigenvalue corresponding to the index and ε is the power-law exponent. We are also going to show that the exponents do not change significantly during 2003 to 2010 which indicates the regularity in the Internet graph at AS level despite of growth in the internet size.

In order to prove the existence of power-laws in our depicted graphs where the eigenvalues of adjacency or normalized Laplacian matrix are depicted versus index, the need of linear regression concept is inevitable. “The attempt to make a mathematical relation between observed variables is known as regression. Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data [27].” “One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable [27].” Linear regression is used to fit a line in a set of two-dimensional points [33][34]. This project uses the least square errors method to plot the linear regression line for the plotted data. “The validity of approximating is indicated by correlation coefficient [33].” A linear regression line has an equation form of Y = b + aX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is a, and b is the intercept (the value of y when x = 0) [27].

Below is an example of regression line y = 1,1091x + 10,709 and R^2 = 0,7142 for a set of data:

index	0	1	2	3	4	5	6	7	8	9
value	9	11	12	14	16	17	16	20	18	19

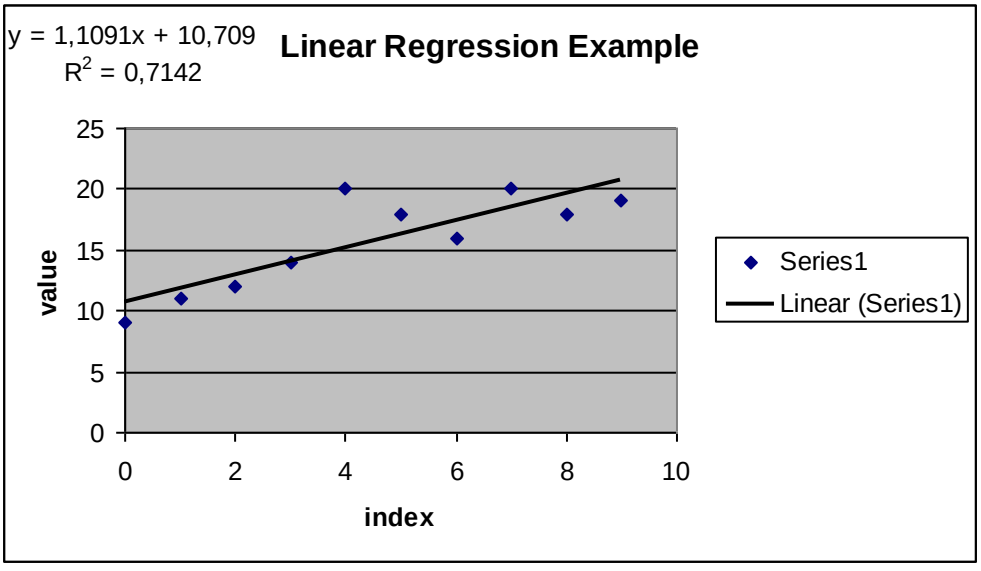

Figure 5. 1: Regression line

There are two important definitions regarding regression line: correlation coefficient and correlation of determination which are used for validity of approximation. They are used as a measure of how well one variable can predict the other and determine the precision you

can assign to a relationship. The more the values are close to the regression line the more precise our prediction will be.

Correlation of Coefficient

"Correlation coefficient or cross-correlation coefficient is an important numerical measurement which various between 1 and -1 and it shows the strength of the association of the observed data for the two variables [27]." If the correlation coefficient is closer to 1 it means that the variables are more linear and the variables fit the best to the linear regression line. Correlation measures the dependability of the relationship (the goodness of fit of the data to that). In mathematic books *linear correlation coefficient* is shown as *r* or *R.* It measures the strength and the direction of a linear relationship between two variables. This project used formula5.2 to calculate the coefficient correlation which is based on least square approximation:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} - \sqrt{n(\sum y^2) - (\sum y)^2}}$$

(Formula5. 2)

Where *n* is the number of pairs of data. For each set of data x and y, a straight line Y = b + aX is drawn such that sum of squares of distances from set of data points to the straight line is minimum.

$$b = \frac{\sum y \sum x^2 - \sum x \sum yx}{n\sum x^2 - (\sum x)^2}$$

(Formula5. 3)

$$a = \frac{n\sum yx - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

(Formula5. 4)

As it was mentioned before r varies between 1 and −1. If *r* is negative, then we say that *x* and *y* has a negative correlation and if *x* and *y* has strong correlation, then r is closer to −1. If *r* is positive, then we say that *x* and *y* has a positive correlation and if *x* and *y* has strong correlation, then r is closer to +1. If there is no linear correlation between *x* and *y* then *r* = 0. If *r* =+1 or *r* =−1 then a perfect correlation exists.

Correlation of Determination

The correlation of determination is shown by r^2 or R^2 also represents the percent of the data that is the closest to the line of best fit which varies between 0 and 1. It is used for prediction which shows the level of certainty in making prediction from a certain graph. If the regression line meets all the variables on scatter plot then it can explain 100% of variation or in other words r^2 =1.

In this Project the existence of power-laws has been proved by the help of linear regression line. In order to find out that the eigenvalues follow the power-laws a linear regression line has been plotted and by the help of correlation coefficient it is proved that the eigenvalues fit the linear regression line in log-log scale by a strong correlation coefficient close to −1.

6. Data mining

As it was mentioned in chapter 3.2, the BGP routing table information has been taken from RIPE project dataset in MRT format. Then MRT format should be converted to human readable format. The process of collecting the BGP routing table information is as below:

Each RRC in different parts of the world collects and stores the BGP routing table information of its own region periodically during specific times in a day. Then RIS collects all the collected information from each RRC in different parts of the world. Two groups of data is stored in each RRC[28]:

- The files in which the file names started with "bview". They are generated every eight hours and they contain the whole BGP table information.
- The files in which the file names started with "update". They are generated every five minutes and they contain all of the BGP packets

This thesis project analyzes the Internet topology on AS level. Therefore, the information about ASs and the way they are connected should be investigated. This kind of information is a part of BGP table information which can be found in bview files. This project collects the BGP routing table information in bview files during the period of 2003.07.30 to 2010.07.30. This project collects the bview files with time stamp of July 30th at 23:59 in each year from 2003 to 2010 from each RRC. All these files are in MRT format. In order to convert them to ASCI files two different tools has been used: zebra_dump_parser in Route Views project and bgpdump from RIPE project. As it was mentioned before, two different tools has been used for converting MRT files so that it is possible to make a comparison between the outcome of both tools. The outcome of the converted files turned out to be the same. Then, all the bview files (which are collected from 17 different RRCs during 2003 to 2010) are converted to human readable text files. Now it is possible to extract the required information from the text files. The information about AS numbers and their immediate neighbors are of our interest since it provides the infrastructure information of ASs and how they are connected to each other. After extracting the required information separately for each year, then the extracted AS paths of each year should be imported to Matlab in order to have an overall view of internet structure at AS level in each year. Since the internet is a huge network with thousands of ASs and connections, the size of text files are very big. For instance without consideration of reserved AS numbers, the largest allocated AS number by the end of 2010.07.30 is 394239. Therefore in order to import the data in to Matlab, it is broken to smaller chunks of data to make it possible for Matlab to handle it. After importing the chunks in Matlab, then it is possible to make adjacency matrix by the help of formula 4.1. Consequently by calculating node degree of adjacency matrix and formula 4.6 it is possible to produce the normalized Laplacian matrix.

7. Spectral Analysis of The Internet Topology

To be able to perform the spectral analysis of the Internet topology, the spectrum of the Internet graph is needed. In order to extract the spectrum of the Internet structure, the corresponding adjacency matrix is needed. This project investigates the spectrum of both adjacency and normalized Laplacian matrix. The adjacency and Normalized laplacian matrix have been created by the help of formula 4.1 and 4.6 respectively. The next step is to calculate the spectrum of the corresponding matrixes. Due to the huge size of the Internet and also to ease the analysis 150 largest eigenvalues of adjacency and normalized Laplacian matrix have been calculated and depicted which will be described further in chapter 7.2.

7.1. Cluster of ASs

This section presents the internet as the cluster of connected ASs during a period of 2003.07.30 to 2010.07.30. The following graphs are based on the adjacency matrix driven from the collected BGP tables in RIPE project during 2003.07.30 to 2010.07.30 where connectivity between two ASs is shown by a blue dot and the empty area between the blue dots indicate disconnectivity. Disconnectivity between ASs have two reasons. The first reason is that the allocated ASs are not connected to each other. The second reason is that, those specific disconnected ASs have not been allocated by IANA during the mentioned time (specified in the caption of each graph). If node 2 is connected to node 6 consequently node 6 is connected to 2 since the adjacency matrix of the Internet topology is symmetric which can also be seen in all the graphs in chapter 7.1. As it was explained before, due to the growth of internet, the need for having more AS numbers eventuated in additional 32-bit format AS numbers. Assigning 32-bit AS numbers has been started in 2006. According to IANA and also the depicted graph in figure 7.1, not considerable 32-bit format AS numbers has been assigned during 2006 to 2010. In general the number of assigned 32-bit format AS numbers from 2006 to 2010 are not considerable. Figure 7.1 and 7.2 are the same plotted graphs of adjacency matrix in 2008 but figure 7.2 excludes the 32-bit format AS numbers. In order to make the analysis easier and more comprehensible it is decided to truncate the original graph (which includes both 16-bit and 32-bit format AS numbers, figure 7.1) to a truncated format (which includes 16-bit format AS numbers, figure 7.2). This truncation has no negative impact on the analysis since the number of allocated 32-bit format AS numbers are not considerable at all. Adding up 32-bit format AS numbers just makes the analysis more complex, since the scale of work will change only because of not considerable amount of assigned 32-bit format AS numbers. Depicting all the 16-bit and 32-bit format AS numbers reduces the clarity and makes it very difficult to extrapolate the results from the graphs. Therefore all the graphs in chapter 7.1 illustrate the 16-bit format AS numbers. So the figures in chapter 7.1 show the growth of the Internet in AS level only in 16-bit format which had the main growth during 2003 and 2010.

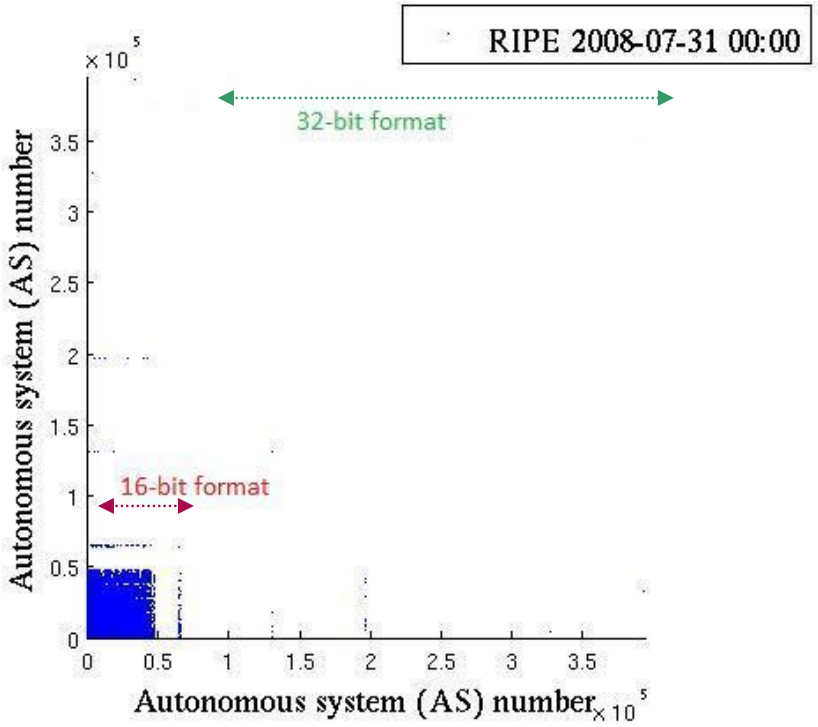

Figure 7. 1: The original adjacency matrix pattern including 16-bit and 32-bit format in year 2008

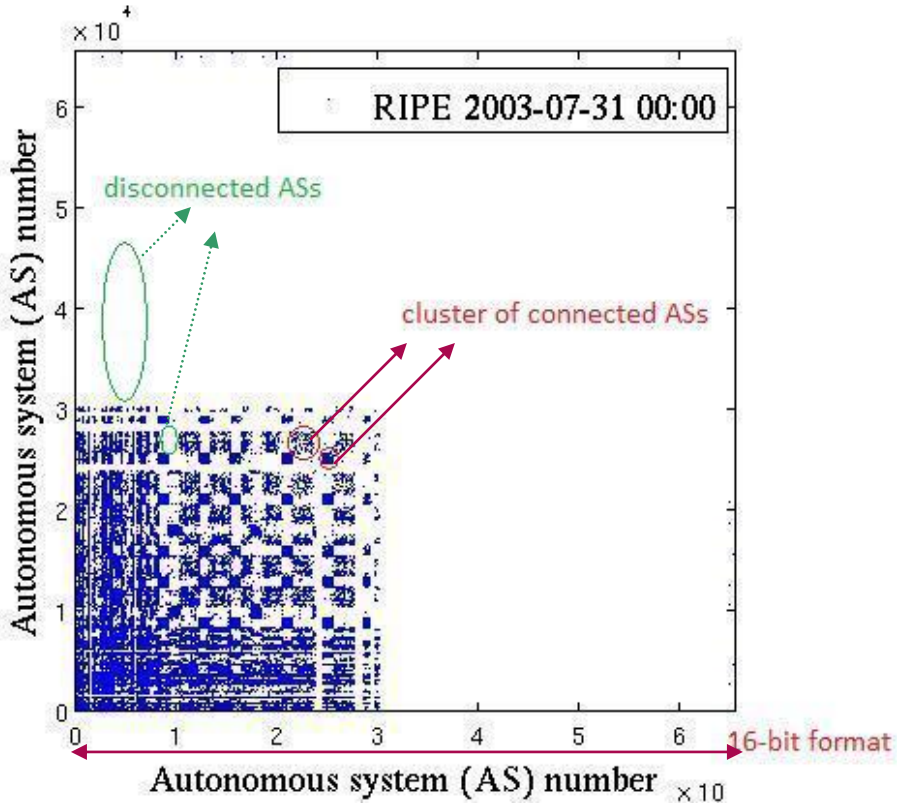

Figure 7. 2: The truncated adjacency matrix pattern including 16-bit format in year 2008

As it can be seen in figure 7.1, there are considerably fewer number of 32-bit AS numbers participating in the original graph. Besides, using the original matrix (figure 7.1) instead of truncated matrix (figure 7.2) makes the comparison of the graphs more difficult consequently it makes the extrapolation and observation more complex in growth of AS numbers perspective. Therefore for ease sake the truncated versions of matrixes will be shown (from 2006 to 2010).

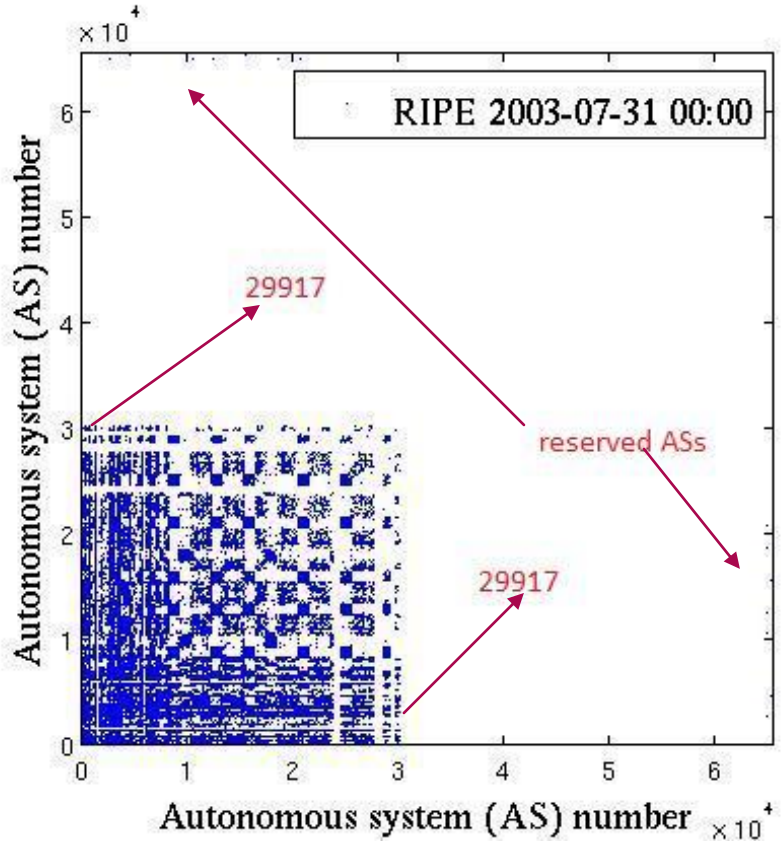

Figure 7. 3: The original adjacency matrix pattern in year 2003

As it can be seen in figure 7.3 the connected ASs are in forms of cluster of ASs. According to IANA (Internet Assigned Numbers Authority) the largest assigned autonomous number [10] by the end of 2003.07.30 is ASN (Autonomous System Number) 29917 which is assigned to ARIN (American Registry for Internet Numbers). It is exactly the same number in our adjacency matrix pattern in figure 7.3. AS it is shown in figure 7.3 there are also some connected nodes in area of 64000 which are the reserved AS numbers for private use. These numbers are also indicated in IANA. (64512-65534 are designed for private use and 65535 is a reserved AS number for special use) [10].

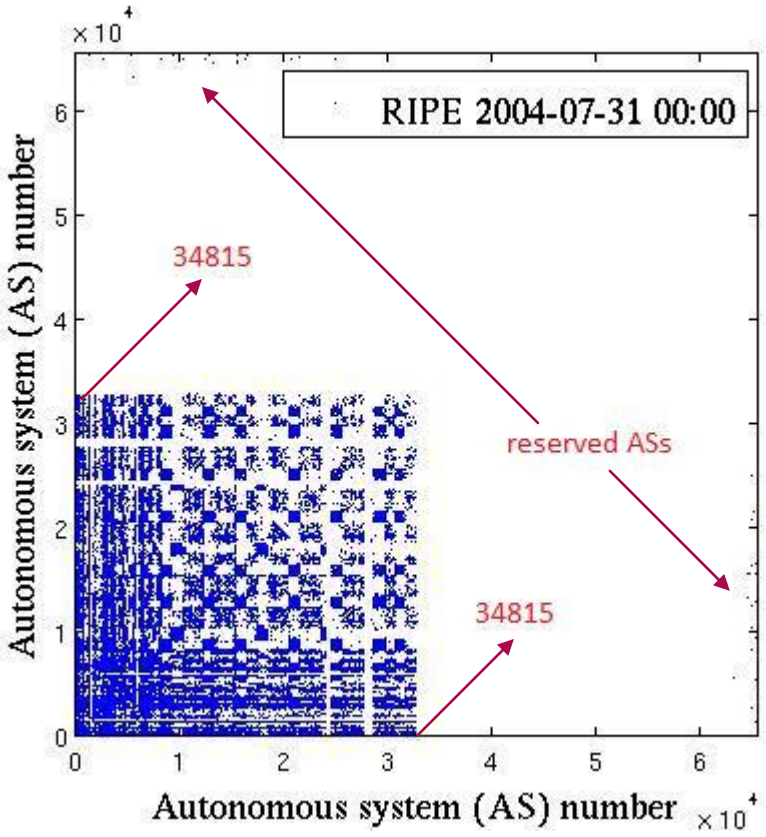

Figure 7. 4: The original adjacency matrix pattern in year 2004

Figure 7.4 illustrates the adjacency matrix pattern in 2004. As it can be seen the pattern and structure of connected ASs are still remained the same as 2003 with clusters of connected ASs. According to IANA by the end of 2004.07.30, the largest assigned autonomous number is 34815 which is assigned to RIPE NCC (The Réseaux IP Européens Network Coordination Centre). It is exactly the same number in our illustrated graph on figure 7.4. Comparison of figure 7.3 and 7.4 illustrates the growth of the Internet in AS level from 29917 to 34815. It can also be seen that more reserved ASs have been used in 2004 than 2003.

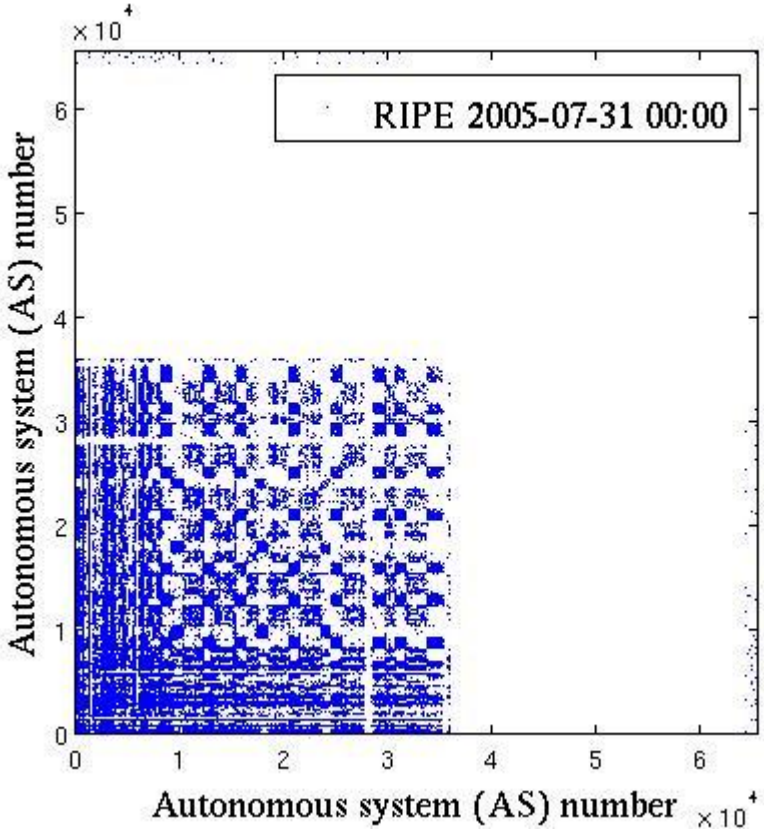

Figure 7. 5: The original adjacency matrix pattern in year 2005

Figure 7.5 shows the adjacency matrix pattern in 2005 which proves that the trend is the same as it was in 2003 and 2004 with clusters of connected ASs. According to information of AS numbers in IANA, there largest assigned number by the end of 2005.07.30 is 38911 which is assigned to APNIC (Asia Pacific Network Information Centre) and can clearly be seen in figure 7.5. The growth of assigned ASs in comparison to 2003 and 2004 can be easily seen. As the figure shows the growth is not limited to the assigned ASs, it can also be seen in the reserved ASs.

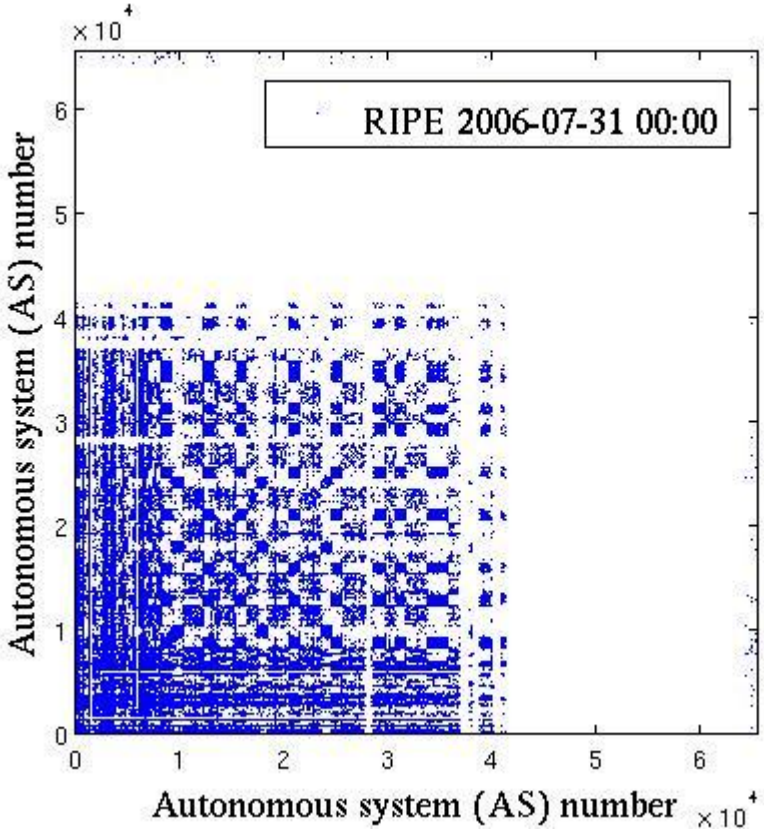

Figure 7. 6: The truncated adjacency matrix pattern only including 16-bit format in year 2006

As it was mentioned before, introducing 32-bit format has been started in 2006. Therefore from 2006 not only 16-bit format was assigned to ASs but also 32-bit format was assigned. Figure 7.6 is a truncated graph which just shows the 16-bit format growth. Comparison between the graphs in 2003, 2004, 2005 and 2006 shows the growth of the Internet in this time period. According to IANA the largest assigned AS number in 16-bit format by the end of 2006.07.30 is 41983 which is assigned to RIPE NCC and it is also illustrated in the extracted graph from the real world data in figure 7.6. Despite of the growth in AS level, the trend still remains the same as the cluster of ASs. According to the graphs the growth in AS level is not similar to random growth; rather it follows a structural growth as the cluster of connected ASs

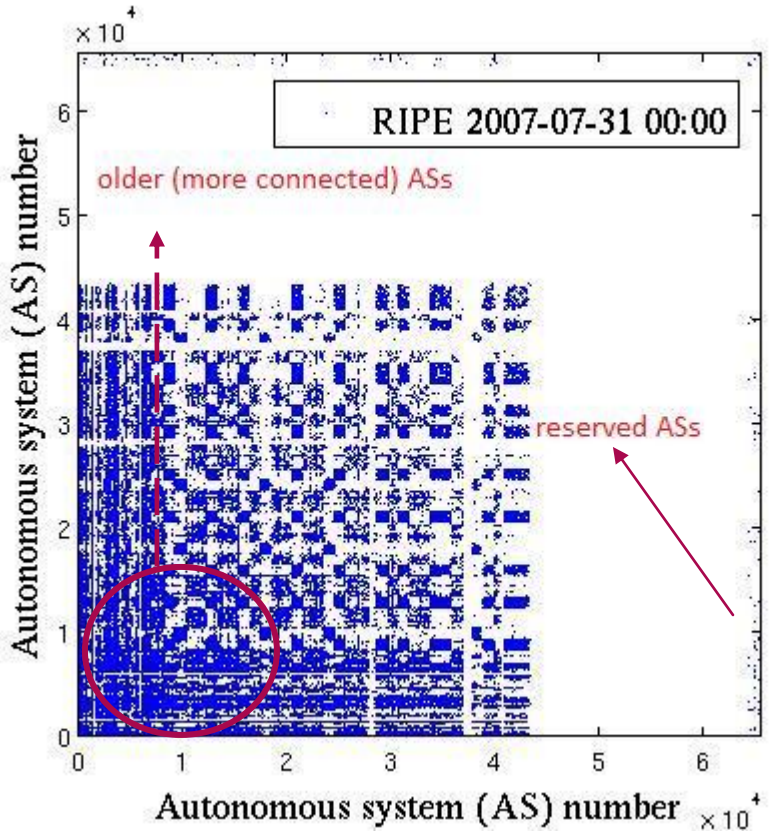

Figure 7. 7: The truncated adjacency matrix pattern only including 16-bit format in year 2007

The graph shows the growth of the Internet in AS level in comparison with previous years. It shows that the older ASs which are more connected and having a higher connectivity are getting even more connected in comparison with other ASs in 2007 than previous years. According to IANA by the end of 2007.07.30 the largest assigned AS number in 16-bit format is 44031 which is assigned to RIPE NCC. Figure 7.7 also shows that 44031 is the largest assigned AS by the end of 2007.07.30. The figure also shows more participation of reserved ASs in 2007 than 2003, 2004, 2005 and 2006.

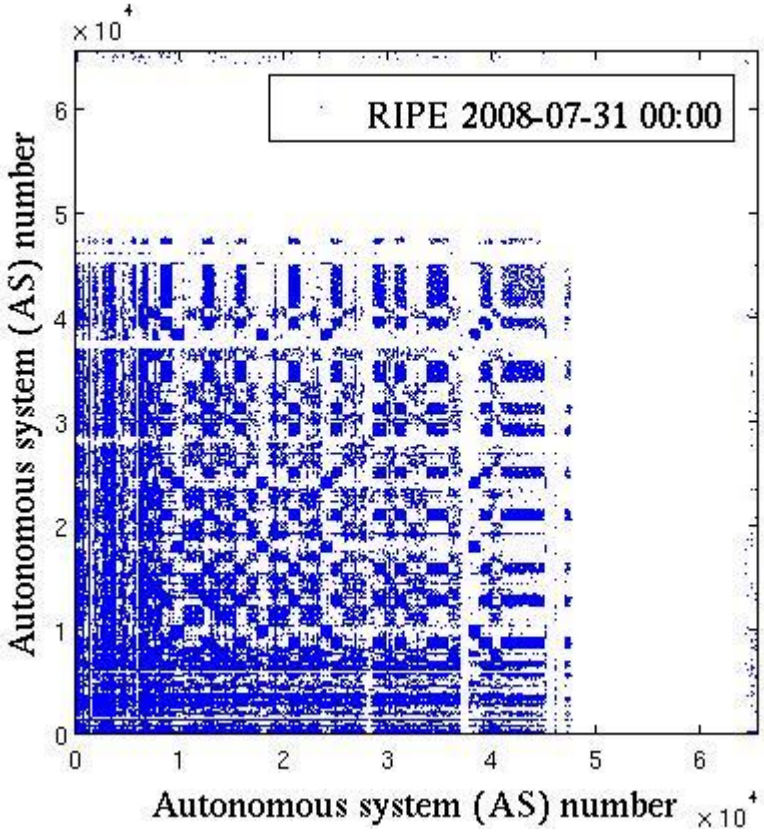

Figure 7. 8: The truncated adjacency matrix pattern only including 16-bit format in year 2008

The picture shows more connectivity in cluster of ASs. The trend of AS growth follows the same structure as the previous years which is the cluster of connected ASs. The older clusters became even more connected in comparison with other ASs than previous years. According to IANA the largest assigned number by the end of 2008.07.30 is 48127 which is assigned to RIPE NCC. The depicted graph from real data also shows that the largest assigned AS number is 48127 by the end of 2008.07.30.

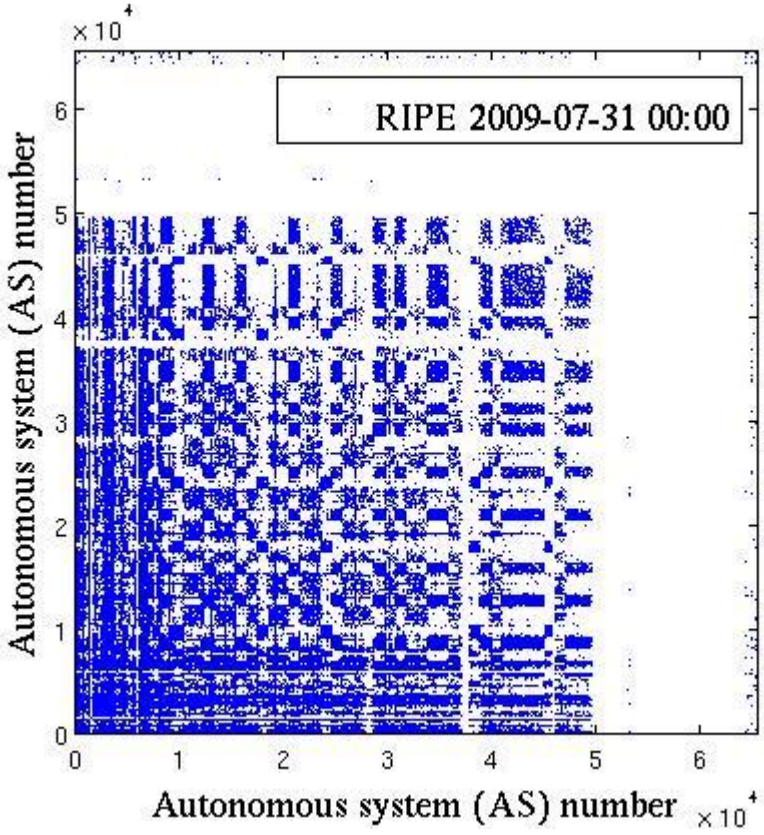

Figure 7. 9: The truncated adjacency matrix pattern only including 16-bit format in year 2009

Figure 7.9 shows the adjacency matrix pattern in 16-bit format in July 2009. It is obvious that the more connected ASs became even more connected but still having the same structural trend which is clusters of connected ASs. It also shows the growth of the Internet in AS level in comparison with last years. According to IANA by the end of July in 2009 the largest assigned AS number is 55295 which is assigned to ARIN and is compatible to the driven graph from real world data in RIPE project in figure 7.9.

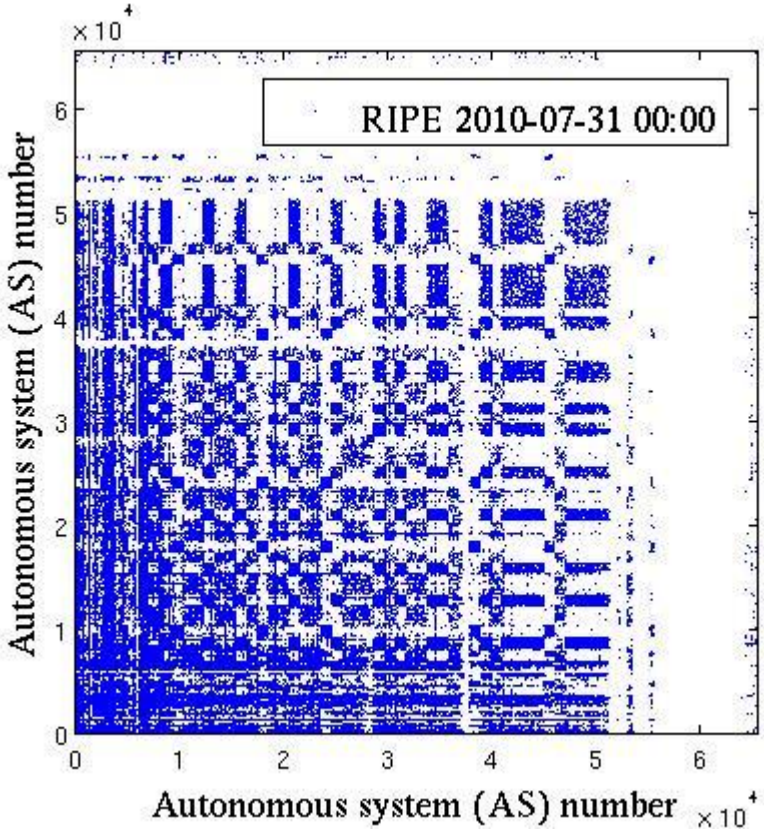

Figure 7. 10: The truncated adjacency matrix pattern only including 16-bit format in year 2010

The picture shows the same trend of clusters of connected ASs and the growth of the Internet in AS level in 16-bit format. More connectivity can be seen in graph 7.10 than all the other graphs in previous years. The comparison of figure 7.3 and figure 7.10 shows that the assigned 16-bit format ASs increased from 29917 to 56319 from 2003.07.30 to 2010.07.30. According to IANA the largest assigned number in 16-bit format by the end of July 2010 is 56319 which is assigned to APNIC and it is also can be demonstrated in our depicted graph from real world data in figure 7.10.

7.2. Eigenvalues & Power law

According to [2][3][4][5] various power-laws exist in the Internet topological properties. Power-laws can be seen in node degree frequency versus node degree, number of nodes within a number of hops versus number of hops, eigenvalues of the adjacency matrix and the normalized Laplacian matrix versus the order of the eigenvalues, and node degree versus node rank (sorting the graph nodes on descending order based on their node degree and indexing a sequence number to them, is called node rank [4]). This sub-chapter shows the existence of power-laws in eigenvalues of both adjacency matrix and normalized Laplacian matrix versus index. In order to observe the existence of power-laws for eigenvalues the 150 largest eigenvalues are plotted in descending order versus indices (ascending sequential numbers) in log-log scale both in adjacency matrix and normalized Laplacian matrix. In order to prove the existence of power-

laws a linear regression line is also plotted. As it was mentioned before the correlation coefficient indicates how well the plotted graph fits to the linear regression line. According to formula5.1 linear regression line indicates the existence of power-laws since it is a linear equation in log-log system which fits the plotted data. Therefore, the high correlation coefficient in the following graphs indicates the more fit of plotted data to the regression line and consequently the existence of power-laws. As it can be seen in the graphs the power-law exponents are calculated from the linear regression lines $10^b x^a$ in log-log scale where slope b is the power-law exponent.

7.2.1. Eigenvalues of Adjacency Matrix & Power law

The following graphs show the 150 largest eigenvalues of adjacency matrix form 2003.07.30 to 2010.07.30 from RIPE project in descending order versus index. It is shown that $\lambda_{Ai} \propto i^{\varepsilon}$ exists

where i is an index, λ_{Ai} is the eigenvalue of the corresponding index therefore ε is the power-law exponent.

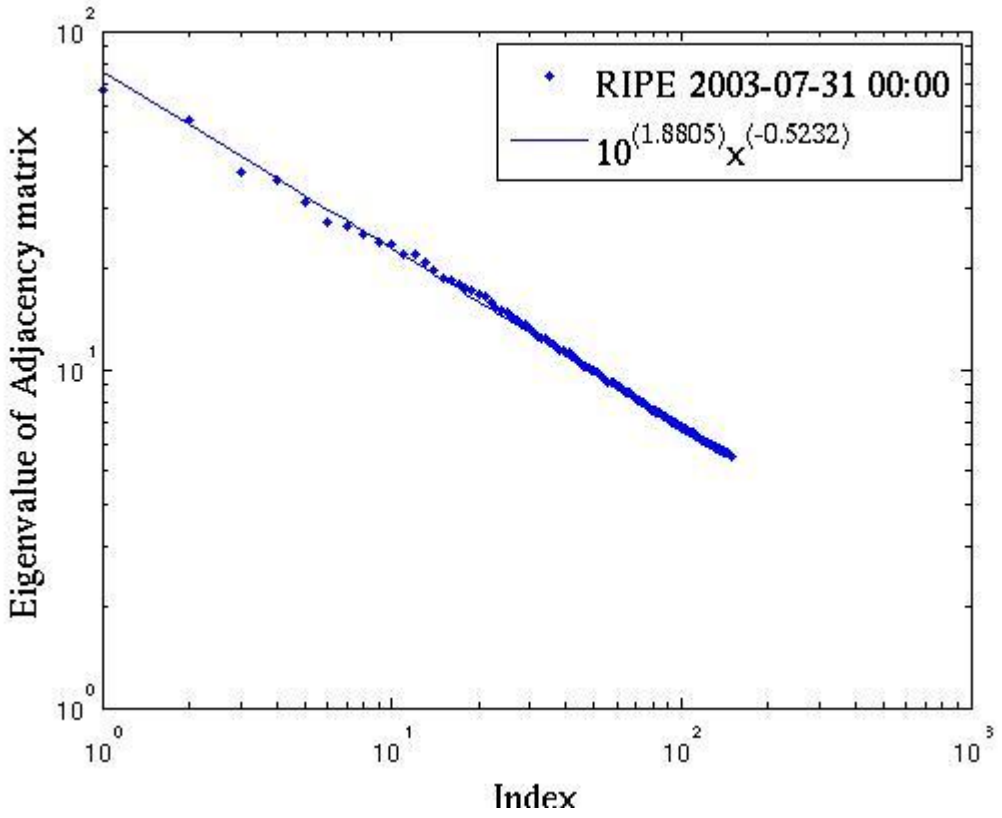

Figure 7. 11: 150 largest eigenvalues of adjacency matrix vs. index in year 2003

As it is shown in the graph power-law exists in eigenvalues of adjacency matrix and the plotted eigenvalues fits the linear regression line with correlation coefficient -0.9989. The power-law exponent is -0.5232.

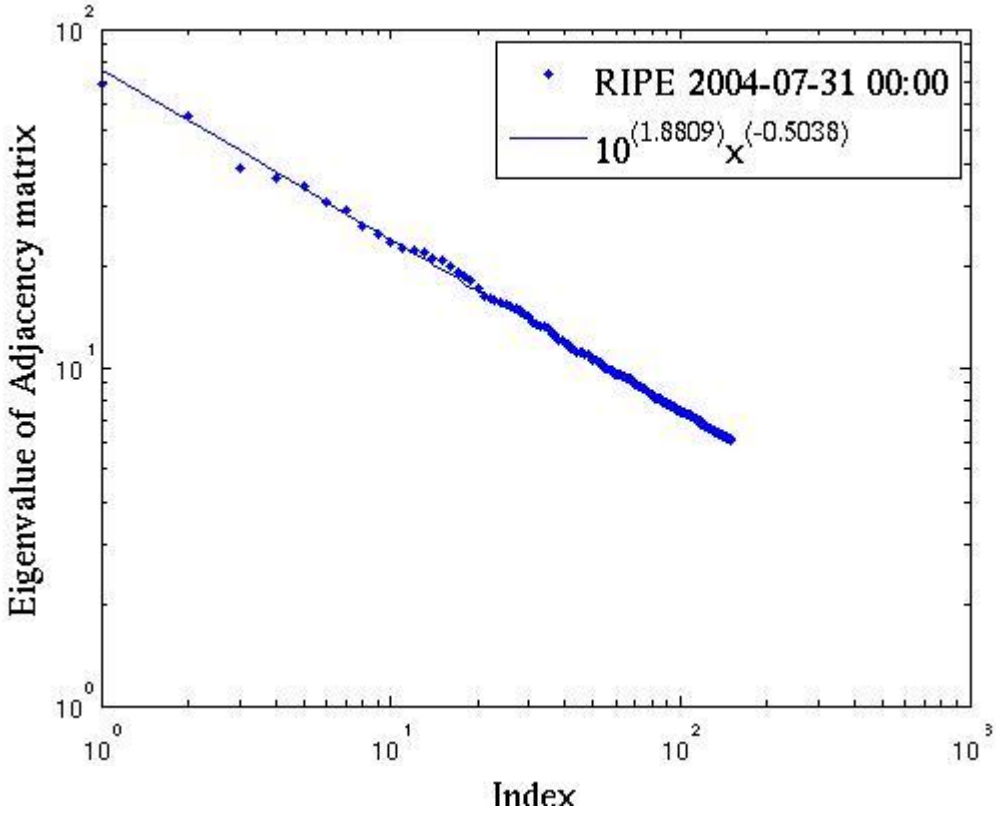

Figure 7. 12: 150 largest eigenvalues of adjacency matrix vs. index in year 2004

As it is shown in the graph power-law exists in eigenvalues of adjacency matrix and the plotted eigenvalues fits the linear regression line with correlation coefficient -0.9991. The power-law exponent is -0.5038. As it is illustrated, despite of the growth and changes in the Internet during one year the power law exists in eigenvalues of adjacency matrix.

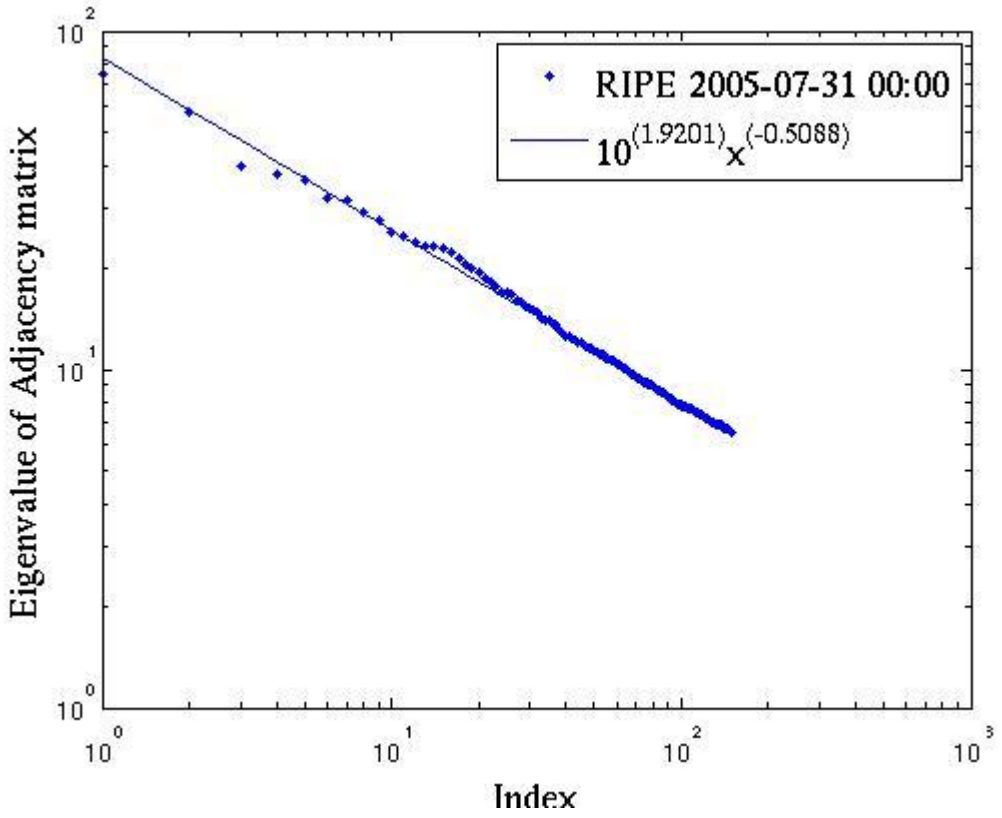

Figure 7. 13: 150 largest eigenvalues of adjacency matrix vs. index in year 2005

Figure 7.13 shows that power law exists in eigenvalues of adjacency matrix regardless of immense changes on the Internet. As it is illustrated there is a high correlation coefficient of -0.9982 with the power-law exponent of -0.5088.

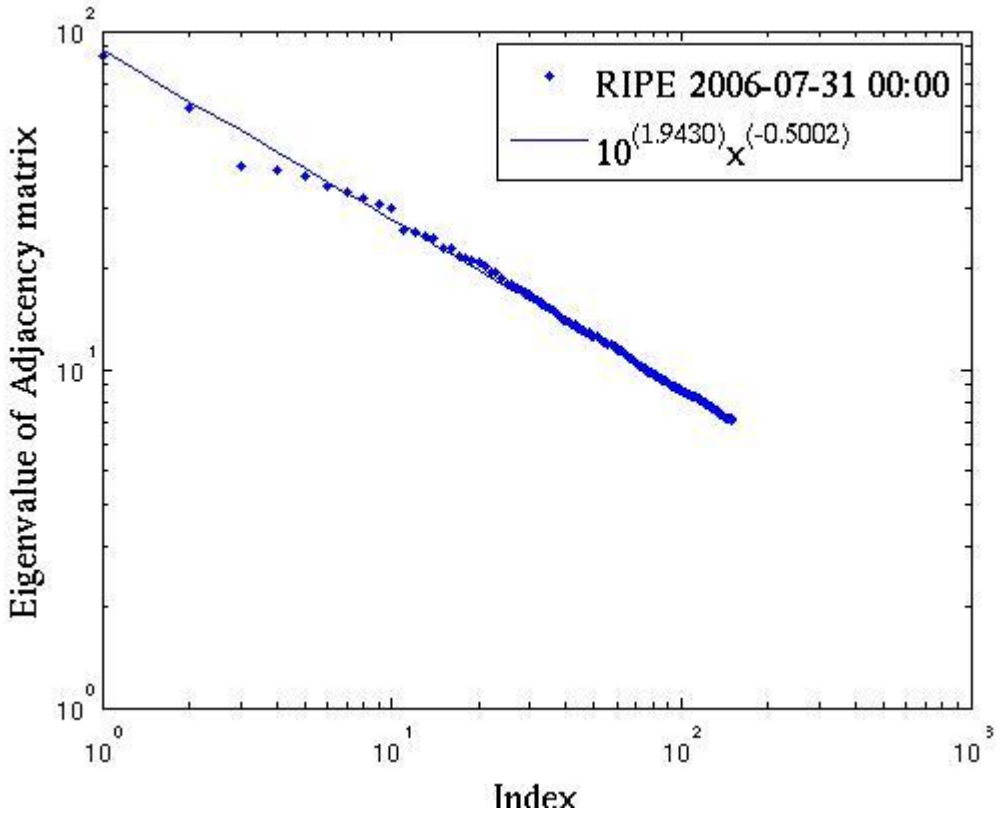

Figure 7. 14: 150 largest eigenvalues of adjacency matrix vs. index in year 2006

The 150 largest eigenvalues of adjacency matrix in 2006 are shown in figure 7.14 which indicates the presence of power-law in eigenvalues of adjacency matrix in 2006. The existence of power-laws can be proven by high correlation coefficient of –0.9980. According to the graph the power-law exponent is -0.5002.

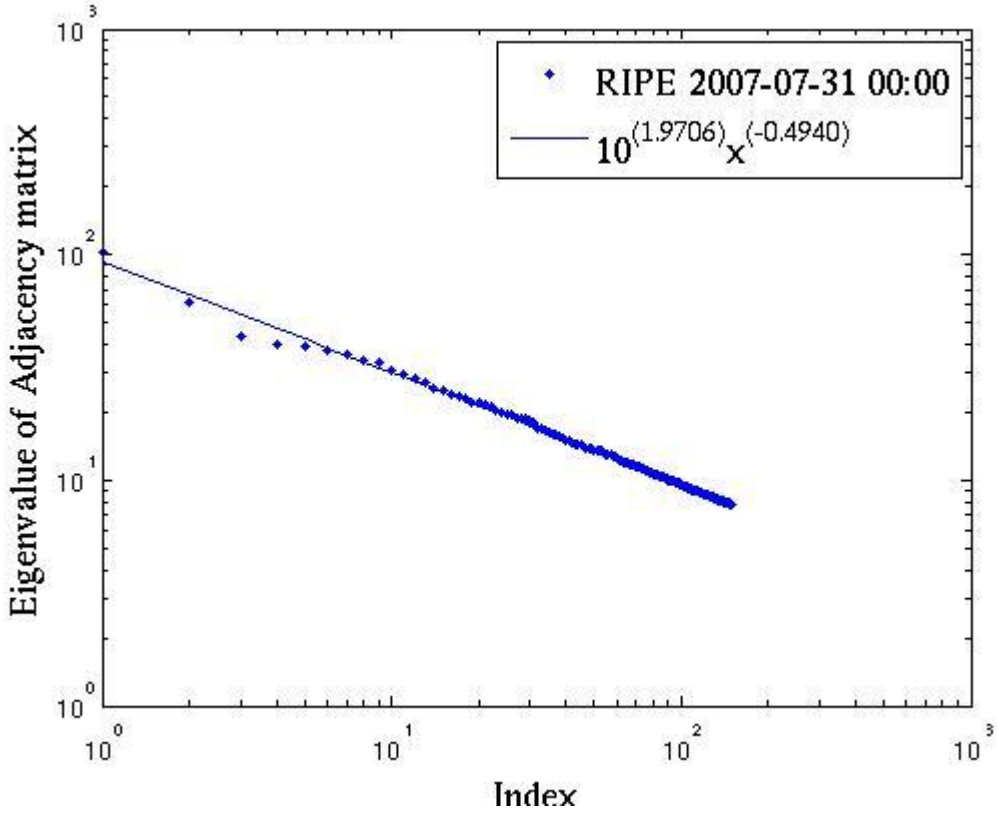

Figure 7. 15: 150 largest eigenvalues of adjacency matrix vs. index in year 2007

Figure 7.15 illustrates the presence of power-laws in eigenvalues of adjacency matrix in 2007 by a high correlation coefficient of -0.9979 and the power-law exponent of -0.4940. It can also be seen, the growth of the Internet results in growth of eigenvalues of adjacency matrix. In comparison with previous years the eigenvalues becomes slightly larger since number of allocated ASs increased. The growth of eigenvalues during time indicates the growth of the Internet size in AS level.

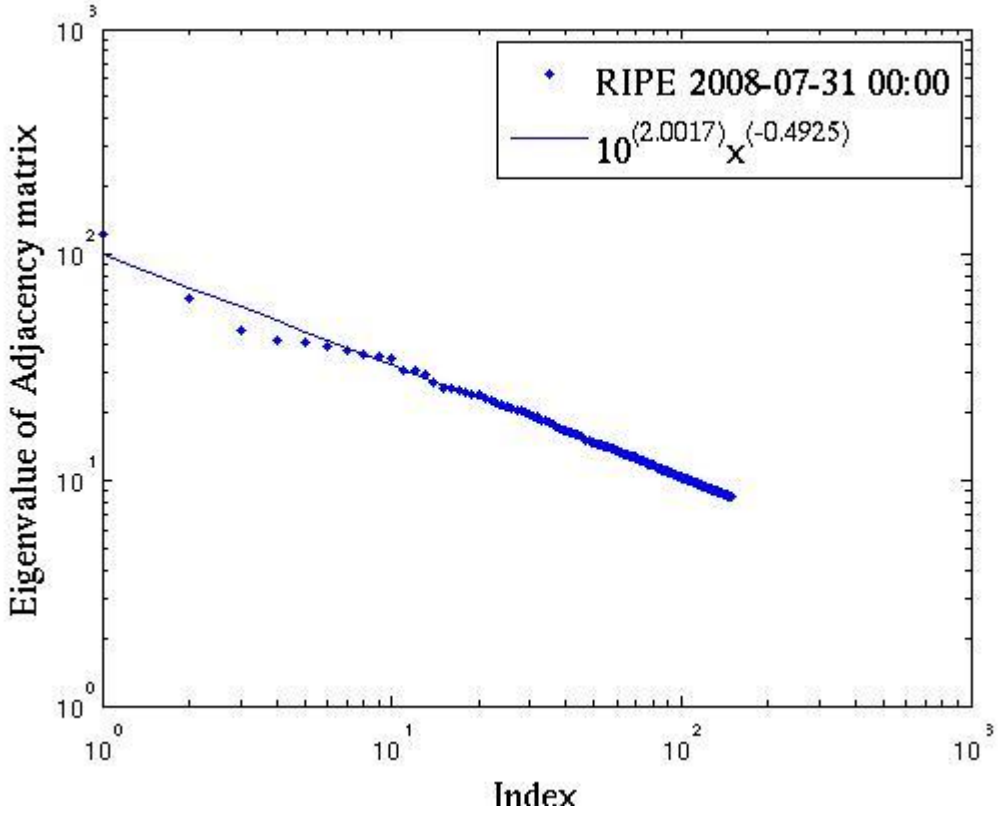

Figure 7. 16: 150 largest eigenvalues of adjacency matrix vs. index in year 2008

The picture proves the existence of power-laws in eigenvalues of adjacency matrix in 2008 by a high correlation coefficient of -0.9969 and a power-law exponent of -0.4925. Despite of growth and changes of the Internet, the existence of power-laws are per persistent. It also shows that by growing of the Internet in AS level perspective, the value of eigenvalues also grows.

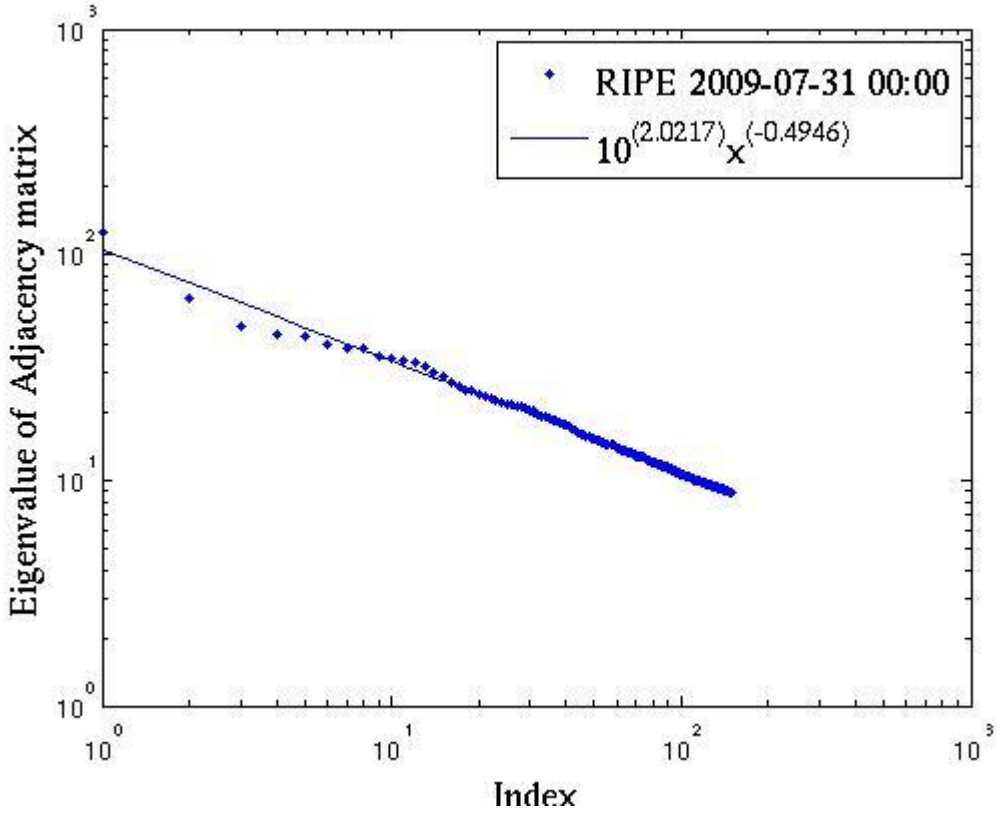

Figure 7. 17: 150 largest eigenvalues of adjacency matrix vs. index in year 2009

The graph for year 2009 similar to the previous graphs shows the presence of power-laws by a high coefficient correlation of -0.9966 and a power-law exponent of -0.4946. It can also be seen that the value of eigenvalues becomes higher in comparison to previous years because of the growth of the Internet in AS level.

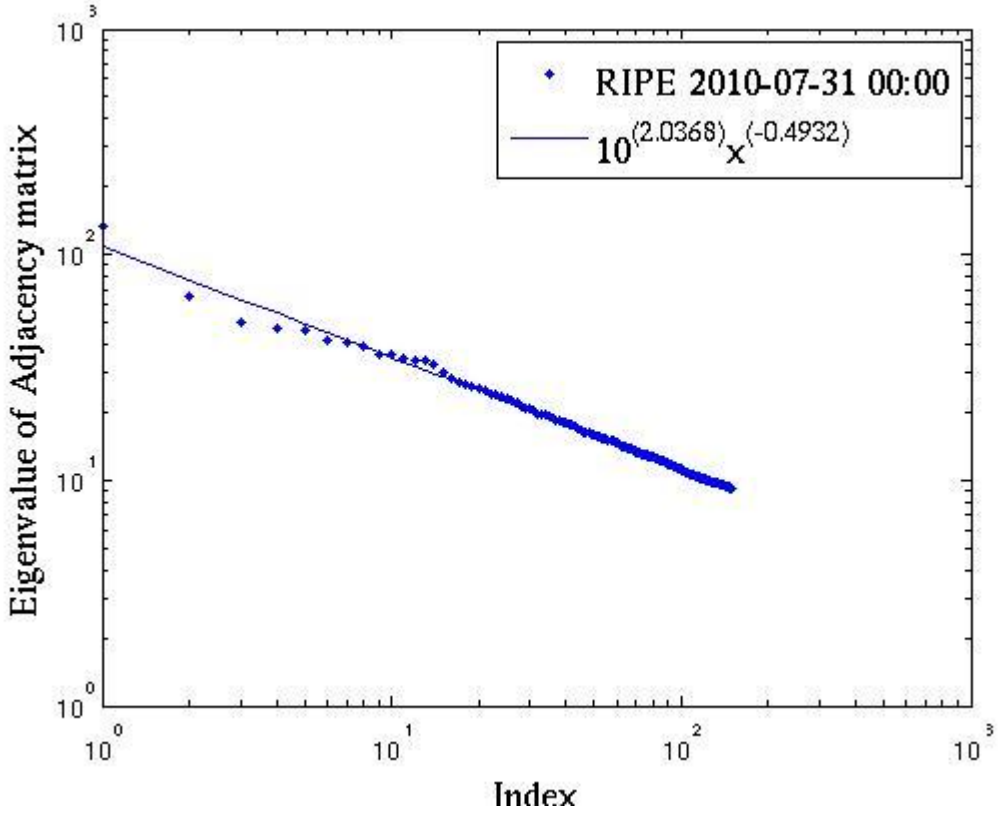

Figure 7. 18: 150 largest eigenvalues of adjacency matrix vs. index in year 2010

The figure illustrates the existence of power-law in eigenvalues of adjacency matrix in year 2010 by a high correlation coefficient of -0.9969 and a power-law exponent of -0.4932. It can also be seen that the values of eigenvalues of adjacency matrix have the largest values between 2003 and 2010 because of the growth of the Internet in AS level.

7.2.2. Eigenvalues of Normalized Laplacian Matrix & Power law

The following graphs show the 150 largest eigenvalues of normalized Laplacian matrix form 2003.07.30 to 2010.07.30 from RIPE project in descending order versus index. We can show that $\lambda_{Li} \propto i^{L}$ exists where i is an index, λ_{Li} is the eigenvalue of the corresponding index, and L is the power-law exponent.

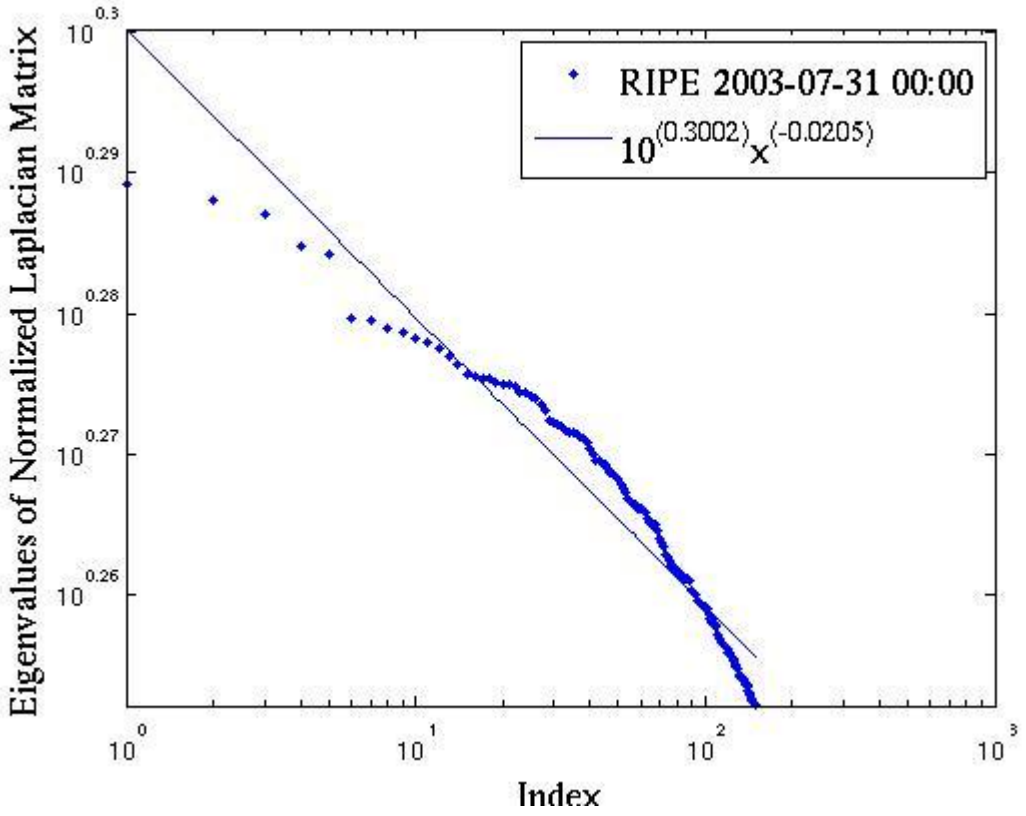

Figure 7. 19: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2003

The graph shows the presence of power-laws in Eigenvalues of normalized Laplacian matrix in 2003. According to the regression line $10^{(0.3002)}x^{(-0.0205)}$, it has a power-law exponent of -0.0205 and a correlation coefficient of -0.9640.

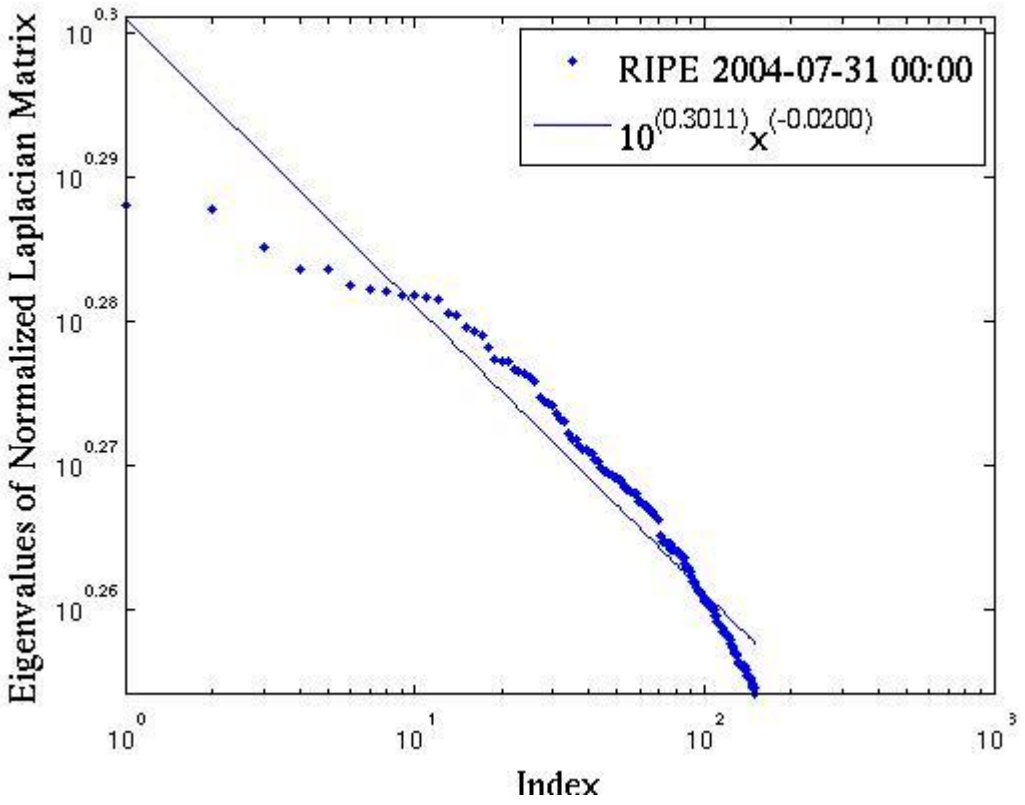

Figure 7. 20: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2004

The depicted graph illustrated the existence of power-law in eigenvalues of normalized Laplacian matrix in 2004. As it can be seen in the picture the plotted graph fits the regression line with correlation coefficient of -0.9626 and the power-law exponent of -0.0200

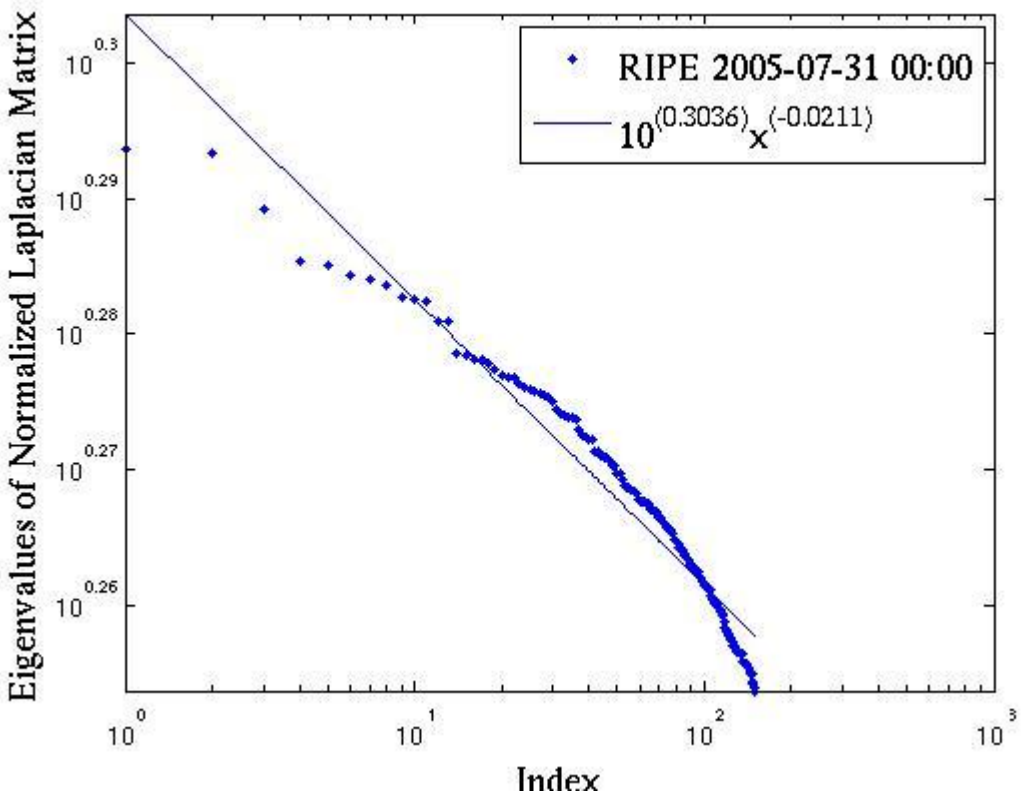

Figure 7. 21: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2005

The plotted graph indicates that despite of growth and changes in the Internet, power-laws exists constantly in eigenvalues of normalized Laplacian matrix. The presence of power-laws can be proven by fitting the plotted data to the regression line $10^{(0.3036)}x^{(-0.0211)}$ with a high correlation coefficient of -0.9709 and the power-law exponent of -0.0211.

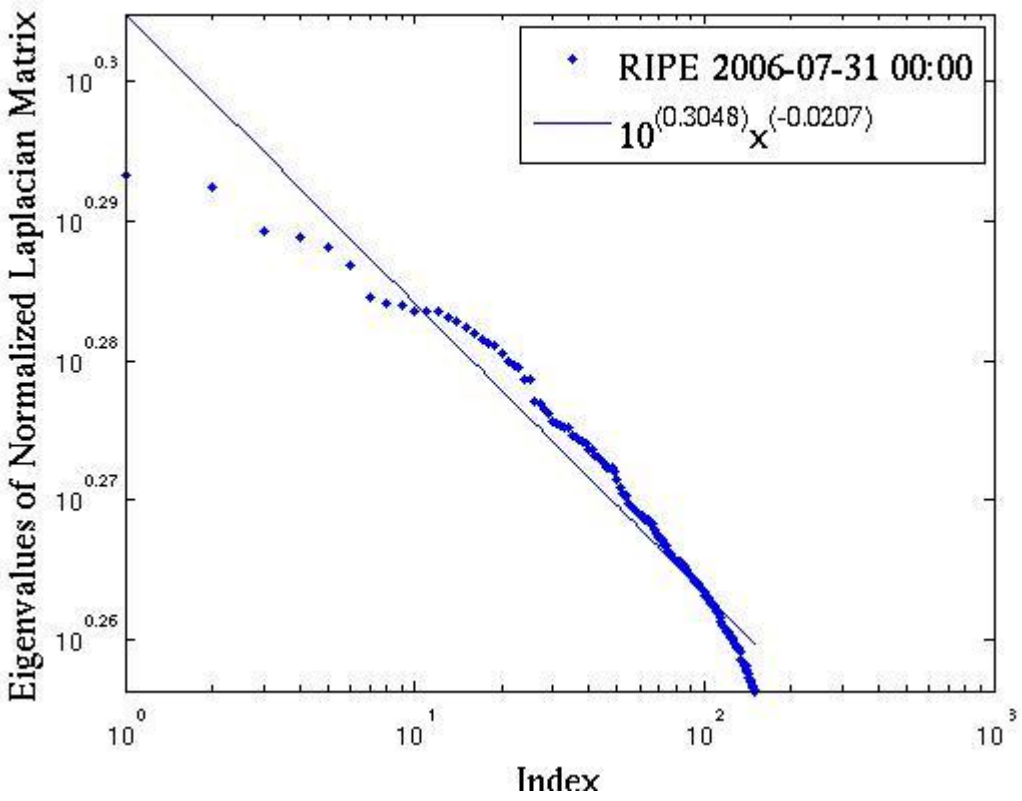

Figure 7. 22: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2006

Figure 7.22 like the previous figures indicates the presence of power-laws in normalized Laplacian matrix driven from real world data by the end of July 2006. The fitting of the plotted data with the regression line is visible in the graph. The coefficient correlation is -0.9731 and the power-law exponent is -0.0207.

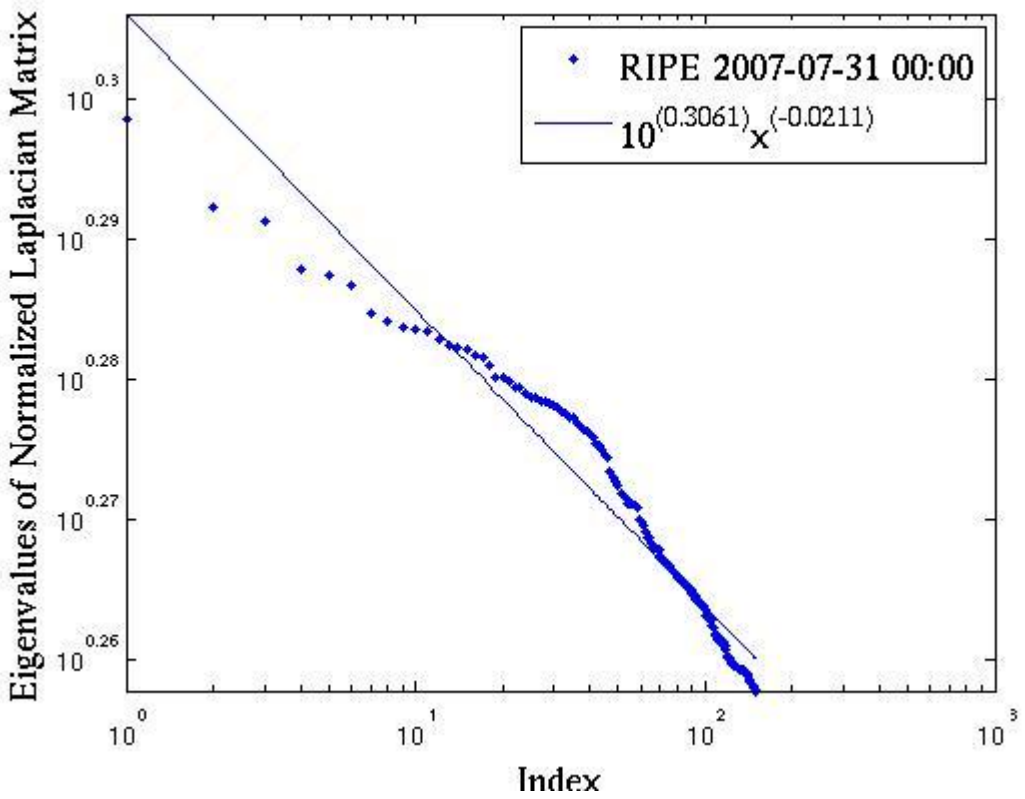

Figure 7. 23: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2007

The depicted graph in figure 7.23 shows the existence of power-laws in eigenvalues of normalized Laplacian matrix in 2007. As it can be seen power-law constantly can be seen in eigenvalues of normalized Laplacian matrix regardless of all changes in the Internet during time. The high correlation coefficient of -0.9693 proves the existence of power-laws. The power-law exponent as it can be seen is -0.0211.

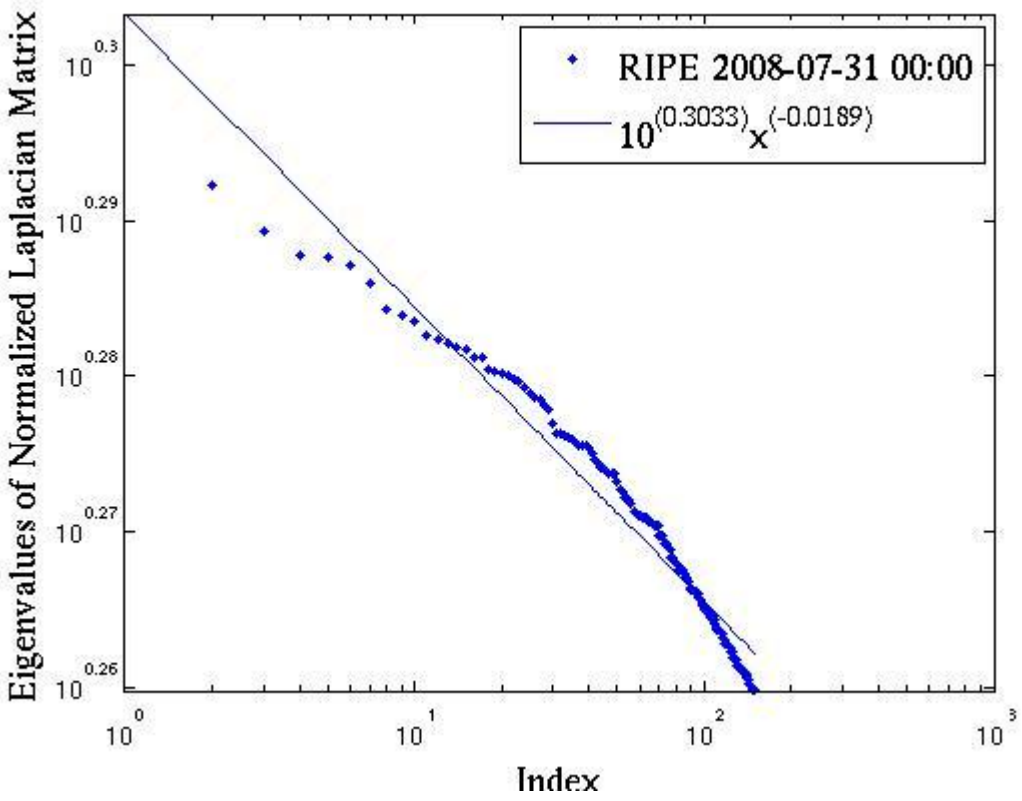

Figure 7. 24: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2008

Figure 7.24 shows the existence of power-laws in eigenvalues of normalized Laplacian matrix driven from dataset RIPE in 2008.07.30. As it can be seen the plotted data fits the regression line with correlation coefficient of -0.9758 and the power-law exponent of -0.0189.

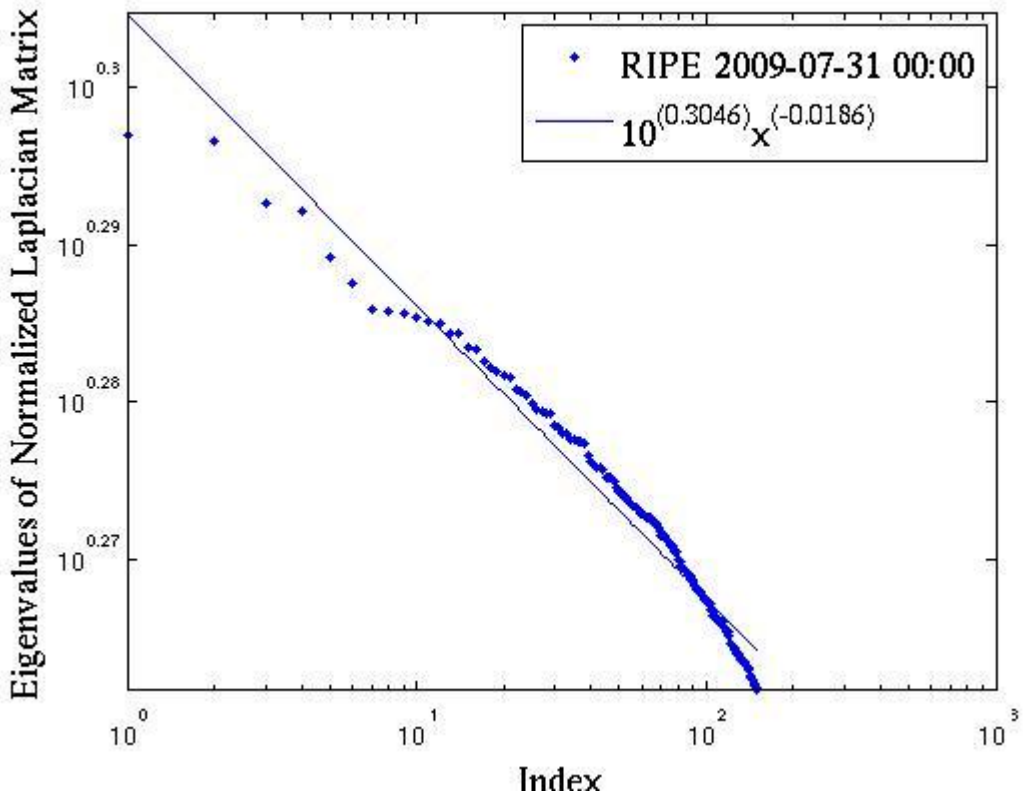

Figure 7. 25: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2009

The picture shows that the eigenvalues of normalized Laplacian matrix driven from real data from RIPE project in 2009 follows the power-laws with a high correlation coefficient of -0.9815 and the power-law exponent of -0.0186. The permanent existence of power-laws in eigenvalues of normalized Laplacian matrix during times can be seen regardless of all the changes in the Internet.

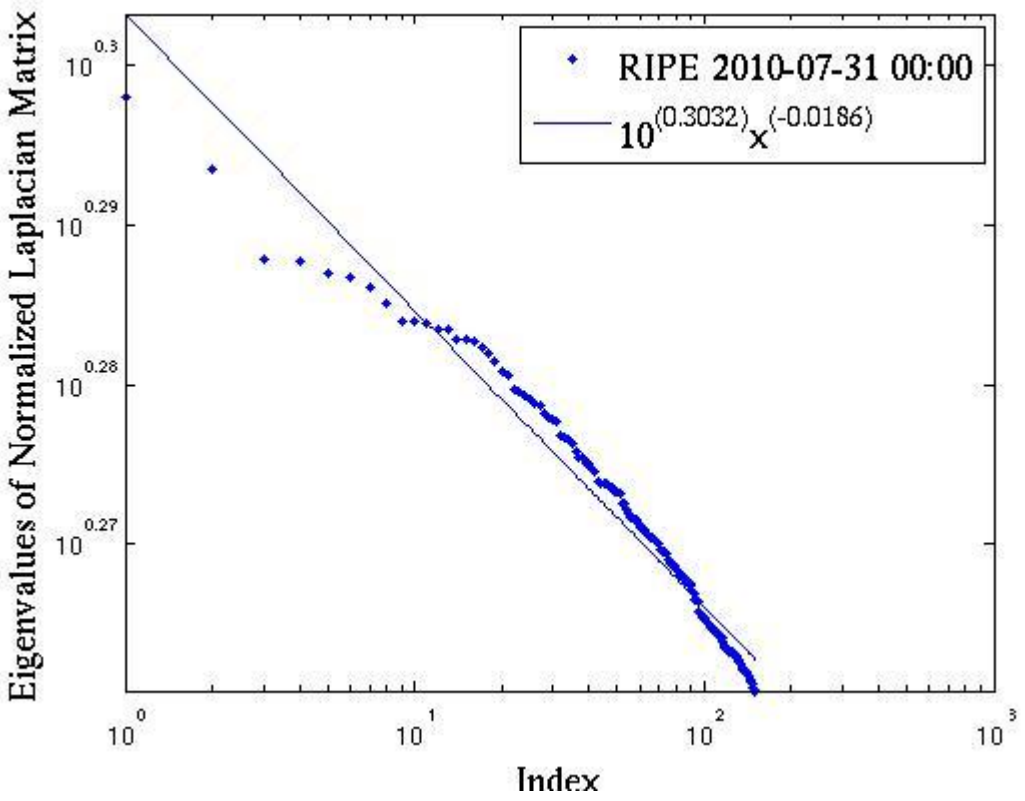

Figure 7. 26: 150 largest eigenvalues of normalized Laplacian matrix vs. index in year 2010

As it was expected from previous graphs, power-law existence can be seen in eigenvalues of normalized Laplacian matrix in 2010. The well-fit plotted data to regression line indicates the presence of high correlation coefficient which is -0.9810. The power-law exponent of -0.0186 is also visible. It can also be seen that during the years 2003 to 2010 power-law exponent values have not changed considerably.

8. Discussion of Results

The purpose of this project is to discover a sort of regularity in a chaotic environment of the Internet. For scientist it is of a great interest to find out how the internet will look like in future. This prediction will help them in more accurate modeling and simulation and consequently the more precise investigation and further studies. As it was mentioned before, this project shows regularities in characteristics of the Internet.

In chapter 7.1 it is shown that the internet grows rapidly during the period of seven years from 2003.07.30 to 2010.07.30 in AS level both in assigned ASs and reserved ASs. An increase in connectivity of ASs can be seen during time. The graphs in chapter 7.1 show that the older ASs have more connectivity and they get even more connected during time. It is also shown that the internet is presented in the form of clusters of connected ASs despite of all irregularities in growth of the Internet size (ex. joining new networks with different protocols of connection) in AS level.

Chapter 7.2 shows regularity in the Internet topology by proving the existence of power-laws in eigenvalues of both adjacency and normalized Laplacian matrix with high correlation coefficient. Power-laws are able to characterize a graph in a single quantity which is the power-law exponent. As it was seen, the power-law exponents have not changed during the seven year of study which indicates regularity in the Internet graph. As the graphs in chapter 7.2.1 demonstrate, the power-law exponents in eigenvalues of adjacency matrix during the period of seven years from 2003 to 2010 did not have a considerable change and remained almost the same with the quantity of 0.5. It is also shown that the power-law exponents in eigenvalues of normalized Laplacian matrix did not change during the period of seven years and it was a constant of approximately 0.2. It is also shown that the eigenvalues became larger during time which is another indication to growth of the Internet in AS level since the number of allocated ASs increased during the period of seven years. Power-law exponents are used for future prediction so it helps the scientists to investigate on different scenarios and find out whether the investigation was correct or not. It proves that the new nodes are not just glued to the network rather they trigger a chain of restructuring change like a fading wave to the rest of the network [33].

Therefore this study can help the people who have the “what if” questions and they want to know how will the Internet look like in the future. It also helps to produce a more accurate and realistic graphs similar to the Internet. There are many different factors involving in dynamic network of the Internet such as economical, cultural, technological, social, and etc. Therefore it is not easy to predict the behaviors of such a network but on the other hand it is possible to characterize the Internet topology by a quantified number which is the power-law exponent.

9. Conclusion

In this thesis real data at AS-level from BGP tables has been analyzed. According to both the theoretical and practical parts of the thesis, some characteristics of the Internet topology follow a certain trend over the time studied.

It has been proved that the Internet graph can be shown as clusters of connected ASs and as the Internet grows in size, it appears the clusters become even more connected. It was shown that during the seven year period, from 2003.07.30 to 2010.07.30 the more connected ASs which are the old ASs became even more connected over time. In fact the trend of growth remaines the same in the form of clusters of connected ASs but the number of connected ASs within a cluster actually increases considerably and even new clusters of connected ASs has been added to the network which indicates the growth of the Internet in size at AS level. It is also proved that the growth in size of the internet is not limited only to the assigned ASs but also the growth goes further to more engagement of private and reserved ASs during the period of seven years.

This thesis has demonstrated that despite of the rapid growth of the Internet without any centralized administration, the power-laws constantly can be seen in eigenvalues of adjacency and normalized Laplacian matrix during the seven year period, 2003.07.30 to 2010.07.30 with high correlation coefficients and there is no considerable change in power-law exponent values during the period which indicates a regularity in the Internet. This thesis has conducted an investigation into eigenvalues of adjacency matrix shown that, the values of eigenvalues become larger over time because the number of allocated ASs are growing at the same time as the Internet connectivity.

10. Reference

[1] D. Morley, Charles S. Parker, "Understanding Computers: Today and Tomorrow, Comprehensive", Boston, MA, United States, 12th edition, 2009

[2] J. Chen and Lj. Trajkovi´c, “Analysis of Internet topology data,” in Proc. IEEE International Symposium on Circuits and Systems, 2004

[3] L. Subedi and Lj. Trajkovic, "Spectral analysis of Internet topology graphs," in Proc. IEEE Int. Symp. Circuits and Systems, 2010

[4] M. Najiminaini, L. Subedi, and Lj. Trajkovic, "Analysis of Internet topologies: a historical view," in Proc. IEEE Int. Symp. Circuits and Systems, 2009

[5] Lj. Trajkovic, "Analysis of Internet Topologies," IEEE Circuits and Systems Magazine, 2010

[6] L. Subedi “Power-laws and Spectral Analysis of the Internet Topology” Master’s thesis in Simon Fraser University2010

[7] Behrouz A Forouzan “TCP/IP Protocol Suite”, McGraw-Hill Publishing Co. 3rd edition,

[8] K. Okumura and Lj. Trajkovic, "Spectral analysis and dynamical behavior of complex networks” in Proc. NOLTA 2010

[9] RFC 1771 http://www.rfc-editor.org/rfc/rfc1771.txt/, 2012.06.08

[10] http://www.iana.org/assignments/as-numbers/as-numbers.xml/, 2012.06.08

[11] http://www.ripe.net/ , 2012.06.08

[12] http://www.ripe.net/data-tools/stats/ris/routing-information-service/, 2012.06.08

[13] http://www.ripe.net/data-tools/projects/faqs/faq-ris/what-is-a-remote-route-collector-rrc/, 2012.06.08

[14] RFC 6396, http://tools.ietf.org/html/draft-ietf-grow-mrt-13/, 2012.06.08

[15] M.Rossi: "MRT dump file manipulation toolkit (MDFMT) - version 0.2", in CAIA Technical Report 090403A, 2009

[16] T. G. Griffin, Z. M. Mao, "Interdomain Routing Streams", Workshop on Management and Processing of Data Streams, Berkeley, June 2003

[17] http://www.routeviews.org/tools.html/, 2012.06.08

[18] M. Scott Cavers, “The Normalized Laplacian Matrix and General Randic Index of Graph”, PhD Thesis, 2010

[19] M. Marcus, H. Minc, "Introduction to Linear Algebra", New York: Dover, 1988.

[20] U. Hamenstädt, "Small eigenvalues of geometrically finite manifolds", Als Ms. vervielfältigt, 2003

[21] D. Easley, J. Kleinberg, "Networks, Crowds, and Markets: Reasoning About a Highly Connected World", Cambridge University Press, 2010

[22] J. Wu, M. Barahona, Y. Tan, & H. Deng, "Robustness of Regular Graphs Based on Natural Connectivity", arXiv:0912.2144v1.

[23] S. Kirkland, Carla S. Oliveira, Claudia M. Justel "On Algebraic Connectivity Augmentation", LINEAR ALGEBRA APPL - Linear Algebra and Its Applications, 2011

[24] M. E. J. Newman, "Power laws, Pareto distributions and Zipf's law", arXiv:cond-mat/0412004v3 2006

[25] http://www.necsi.edu/guide/concepts/powerlaw.html/ , 2012.06.08

[26] H.E Stanley, L.A.N Amaral, P Gopikrishnan, P.Ch Ivanov, T.H Keitt, V Plerou, "Scale invariance and universality: organizing principles in complex systems", Physica A:Statistical Mechanics and its Applications, Volume 281, Issues 1–4, 15 June 2000, Pages 60-68

[27] http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm/ , 2012.06.08

[28] http://www.ripe.net/data-tools/stats/ris/ris-raw-data/ , 2012.06.08

[29] Frank J. Hall, “The Adjacency Matrix, Standard Laplacian, and Normalized Laplacian, and Some Eigenvalue Interlacing Results”, Department of Mathematics and Statistics, Georgia State University, Atlanta, 2010

[30] Fan R. K. Chung “Spectral Graph Theory”, Conference Board of the Mathematical Sciences, 1997

[31] B. Auffarth “Spectral Graph Clustering”, course report for T´ecnicas Avanzadas de Aprendizaje at Universitat Politecnica de Catalunya, 2007

[32] satu e. schaeffer “Graph Clustering”, SienceDirect, 2007

[33] M. Faloutsos, P. Faloutsos, C. Faloutsos“On Power-Law Relationships of the Internet Topology”, ACM New York, NY, USA, 1999

[34] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery, “Numerical Recipes in C.” Cambridge University Press, seconnd edition, 1992.