INTRODUCTION TO THEORY OF SCALE FREE NETWORK AND ITS APPLICATION ON CVUT.CZ DOMAIN

Jana Broncová, Tomáš Liška
Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering,
Department of Software Engineering, Trojanova 13, 120 00 Prague 2
broncjan@fjfi.cvut.cz, tomas.liska@cern.ch

ABSTRACT:
The theory of scale free networks was developed to model and to describe real networks. The real networks are networks of our real world. The examples are social networks, nervous systems, living organisms or the WWW net. The first step for understanding was the finding of self-organized character in these networks and their continual development. This article starts with the history of seeking for a model description of the real networks to the discovery of scale free model. It introduces the theory of scale free networks on the example of the WWW net from the data collection and analyses to the verification of some of the typical characteristics. The data source is a network of documents on the cvut.cz domain.

KEYWORDS:
Scale-free network, BA model, topology of WWW.

1 Introduction to Theory

1.1 Predecessor of Scale-Free Networks

Random Model

Networks of various systems are created by very different rules. That is why two famous mathematicians Erdösy and Rényi decided to ignore this variety and join the nodes in network randomly. This theory of random graphs was first published in 1959.

At the beginning there are isolated nodes. The edges are randomly added to nodes in sequence. Some of nodes gain more edges than others; some of them are without an edge for some time. But it was proved that all the nodes have very similar number of edges finally and variances from average value are extremely rare.

This theory affected the next research on networks. If some network was too complicated, it was described as random. But even the society, cells, communication networks or the economy are complicated, these real networks have some rules and they can not be organized by the same principles as the random networks.

Clusters

In the 90s, two scientists Watts and Strogatz examined the networks describing people's influence among themselves. They defined the coefficient of clustering. This coefficient describes the probability of interaction among subjects. It is defined as the quotient of the real number of links among the nodes in graph and the links in the case of the complete graph. The coefficient can gain the values from interval 0 to 1. The value about 0 means the independence among subjects, on the contrary the value about 1 means big influence among subjects.

The proof was made on the small social network where all the relations were known. This network was made by the group of 70 000 mathematicians (nodes) and their shared responsibility in articles (edges). The conclusion was that the coefficient of clustering of this network is approximately 10 000 times greater than the coefficient of random network should be.
Strogatz and Watts decided to make an alternative model to the random model: the nodes are placed in circle and linked to their closest neighbours. In addiction some edges connects the long-distance nodes. These few edges reduce very dramatically the average length between two random nodes.
The difference between Erdösy – Rényi model and Strogatz – Watts model is only the initiatory placing of nodes, so both of them are random. It was necessary to find a model which relations are not created by hazard.

Differences between Real and Random Network
There are two main discrepancies between real networks and previous models:
1. Dynamic vs. static number of nodes: random models start with a fixed number of nodes N which are constant in time, in contrast real networks are open to adding new nodes to the system.
2. Preferential attachment: the probability of linking of two nodes is random and uniform for random models. In fact the adding a new node to the old nodes is preferential in real networks. There is a higher probability to be linked to a node that already has a large number of connections.

1.2 Models of Scale-Free Network

Basic BA Model
This simple model is known as BA model (named after Barabási and Albert). It was published in 1999. This model is also known like "rich get richer" model. It is defined in two steps [5]:
1. Growth: Starting with a small number (m_0) of nodes, at every time step we add a new node with $m(\leq m_0)$ edges (that will be connected to the nodes already present in the system).
2. Preferential attachment: When choosing the nodes to which the new node connects, we assume that the probability Π that a new node will be connected to node i depends on the connectivity k_i of that node, such that $\Pi(k_i) = k_i / \sum k_j$.
After t time steps the model leads to a random network with $N=t+m_0$ nodes and mt edges. The probability that a node has k edges following a power law ($P(k) \sim k^{-\gamma}$) with an exponent $\gamma_{model}=2.9 \pm 0.1$. The scaling exponent is independent of m, the only parameter in the model. $P(k)$ is independent of time and the system size ($N=t+m_0$). It indicates that despite its continuous growth, the system organizes itself into a scale-free stationary state.

Modified Models
There are many models modified from BA model. Some of the examples are Fitness model [3] (combination of a fitness metric and the preferential attachment) and Pennoc's model [7] (combination of a preferential and uniform attachment).

1.3 Characteristics of Scale-Free Networks on Topology of WWW

WWW network is a large directed graph, whose nodes are documents and edges are the links (URLs) pointing from one document to another. BA model noticed in last paragraphs is not able to fully describe the topology of WWW. For example the links are not invariant in time. They are eliminated or rewired to other documents. Similarly, documents are not stable, they are deleted or their address is changed. Moreover web pages are structured in domains and

their structure is more hierarchical. For more realistic model is necessary to consult these items too.

Application of BA model on WWW:

1. The crawler adds URLs found on a document to the database and recursively follows these to retrieve the related documents and URLs.
2. The probability that a document (node) has *k* incoming/outgoing links follows the power law.
3. Power law indicates that the probability of finding the document is higher with a large number of links.

Centers

Scale-free network is composed by abnormally interconnected nodes called centers *(*Figure 1*).* For example the very famous web page has many incoming links, so the probability of finding this page is higher.

The number of incoming and outgoing links is random and not limited on web pages. These are the characteristics of very interactive self-organized systems.

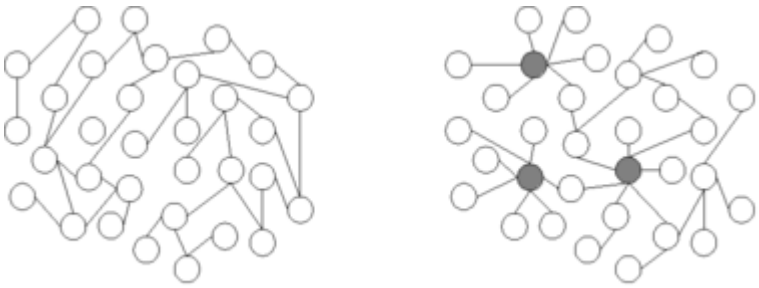


Figure 1 – Random and Scale-Free Networks (black nodes are centers)

The Shortest Path between Two Nodes

The shortest path between two documents, *l*, is defined as the smallest number of URL links one needs to follow to navigate from one document to the other. The average of *l* over all pairs of nodes follows $l=0.35+2.06log(N)$. For a given *N, l* follows a Gaussian distribution, *l* can be interpreted as the diameter of the web, it measures the distance between any two nodes in the system.

It was proved that independent power graph with γ in interval (2;3) has very small diameter $l \sim lnlnN$. So the diameter of scale free network can be regarded as constant.

Robustness – Random Failure Resistance

By robustness we mean the random failure resistance. The robustness is given by number of alternative paths among majority of pair of nodes. The elimination of few nodes has no the impact to the integrity of network, but if the number of eliminated nodes gains the critical value, the system disintegrates. This critical value for scale-free networks is when the $\gamma<3$. So for WWW it means that the disintegration of network happens after elimination of all nodes (so it means never). This topological stability is given by centers. These centers carry the whole network in one piece. The random failure does not recognize the nodes (small node or center) and the probability of failure is the same. There are many small nodes in the system, but only few of centers, so in many cases this random failure happens to any of the small nodes. Even the elimination of few centers is not a problem. Scale-free networks are random failure resistant.

Goal-directed Attack Resistance
If the elimination of nodes is not random, but someone is attacking the centers, the scale-free network disintegrates very quickly. By the elimination of 5-15% of centers the whole network disintegrates to the fragments, this is called the critical value. Even the elimination of 2% of the biggest nodes affects the decreasing of an efficiency of the network by 50%. The scale-free networks are very vulnerable to directed attacks *(*Figure 2*)* and *(*Figure 3*).*

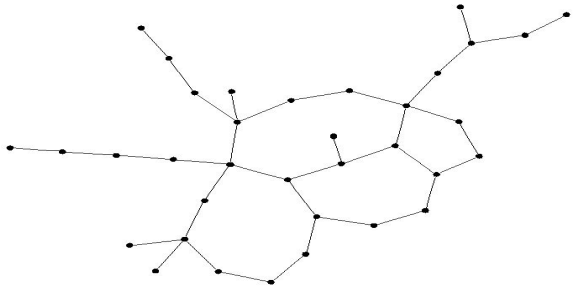


Figure 2 – The original network with 4 centers

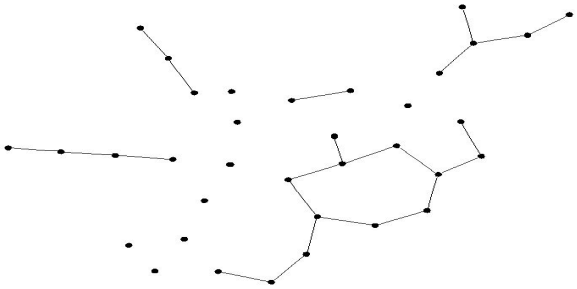


Figure 3 – Eliminating of centers leads to disintegration of the system

2 Data Collection

We mapped cvut.cz domain and gained the data which we need for verification the scale-free characteristic and some more characteristics (centers, resistance) of WWW.
We constructed a simple crawler. This crawler added URLs (only cvut.cz) found on a document in cvut.cz domain, it continued recursively.

2.1 ER Model and Algorithm

The main part of ER model *(*Figure 4*)* is composed by four tables WEBSERVER, HOSTNAME, WEBPAGE and LINK. Data from these tables were used for data analyses. Tables HEADERS_CACHE and PAGE_CACHE served for a temporary data storage in phase of data collection and script debugging.

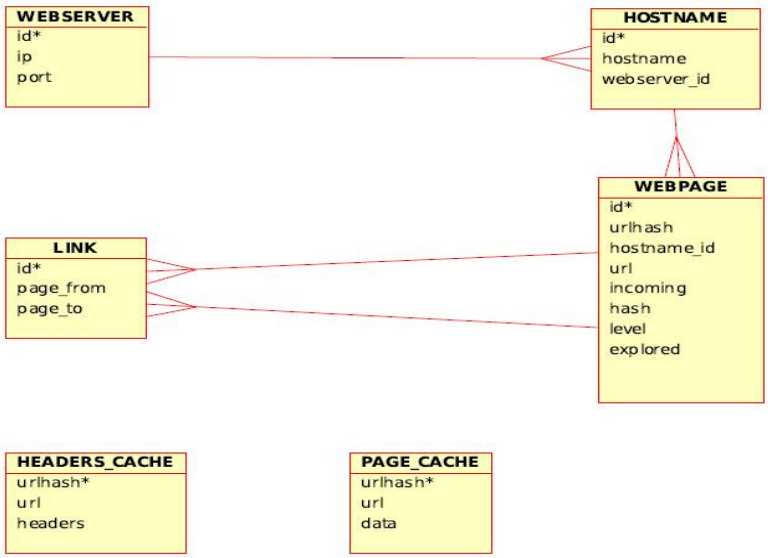
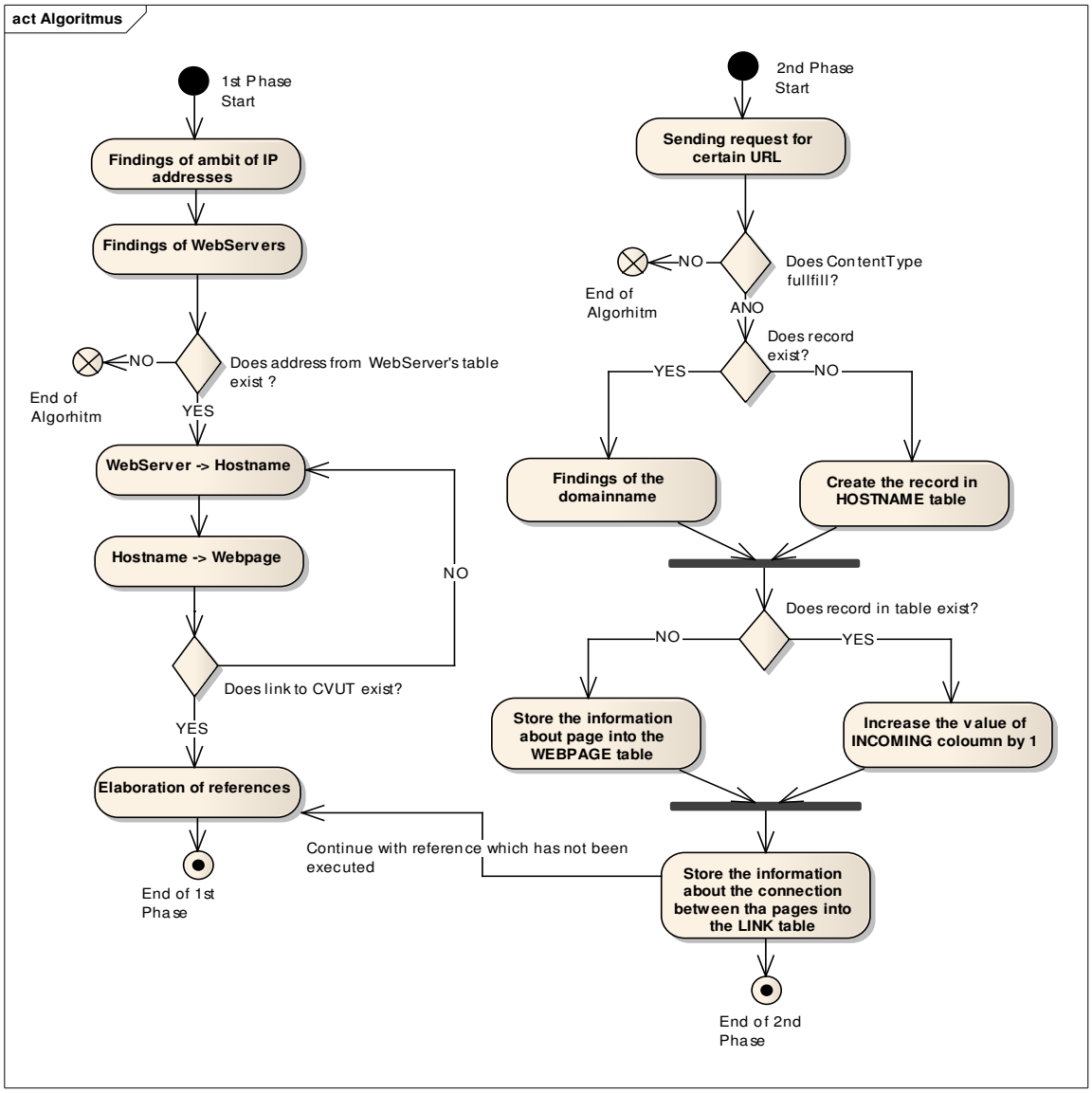


Figure 4 – ER model

Description of ER model:
- WEBSERVER contains the list of all web servers from IP 147.32.0.0 - 147.32.255.255.
- HOSTNAME contains the list of gained host names in IP address form or domain name and foreign key to table WEBSERVER.
- WEBPAGE contains information about web pages
- LINK contains information about interconnection of web pages
- HEADERS_CACHE have information about headers of web pages. This table serves for faster data processing.
- PAGE_CACHE helps in testing data collections.



2.2 Collected Data

The collected data were visualized *(*Figure 5*)* by the tool Neato, which is a part of the package 'graphviz' from the Debian Linux distribution. We found 1337 web servers on port 80 with IP address 147.32.x.x. Other 21 web servers were found by the process. 374 928 unique web pages were gained.

Table	*Number of items*	*Size*
HEADERS_CACHE	3 798 344	2.9GB
HOSTNAME	2 293	126.7kB
LINK	7 411 401	286.1 MB
PAGE_CACHE	2 355 242	27.8GB
WEBPAGE	374 928	590.8MB
WEBSERVER	1 433	96.9kB

Summary – collected data

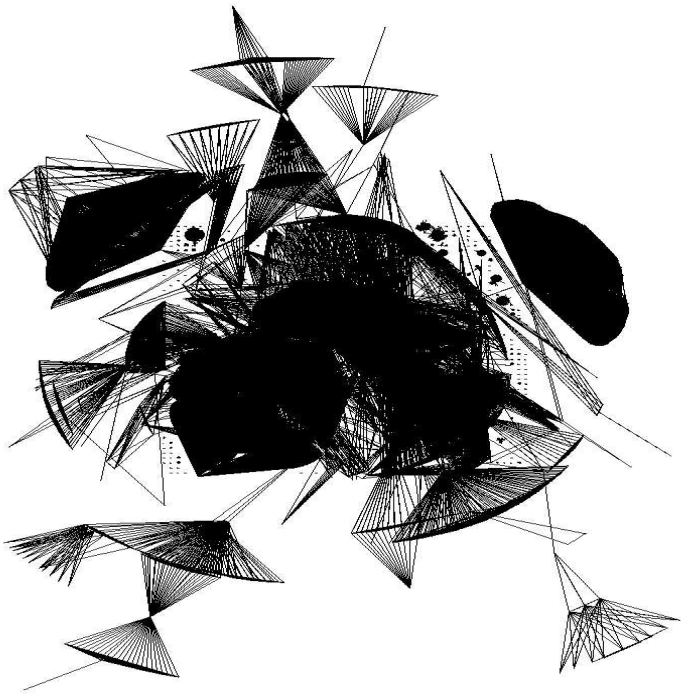


Figure 5 – Visualization of cvut.cz network. Dark areas are highly interconnected nodes.

3 Data Analysis

For data analysis we used the BA model (power law distribution and preferential attachment).

3.1 BA Model

The BA model is based on two statements *(*1.2 Models of Scale-Free Network*).* The first condition is the adding of a new node in every time step. For WWW we can confirm this intuitively. The second condition is the existence of power law in collected data.

Analysed Data

Theoretical prediction are $\gamma=2.9\pm0.1$ for probability $P_{out}(k)$ and $P_{in}(k)$ by a sequence connecting or γ =ln3/n2 by parallel connecting. So it means that the theoretical prediction depends on a precise definition of a model. For the confrontation we used the experimentally measured values by Barabási and Albert ($\gamma_{out}=2.72$ a $\gamma_{in}=2.1$). Data were analysed by the software GRETL (http://gretl.sourceforge.net/), open-source software for statistic analysis.
The indexes came out $\gamma_{in}=1.04$ and $\gamma_{out}=2.39$ *(*Figure 6*)* and *(*Figure 7*).* For outgoing links the result was very similar to their value, for incoming the value is different. There are 841 different values of number of edges for incoming links in total - from 0 (the node without an edge) to 111 916 from all nodes (374 928). For outgoing links there are only 280 different values of number of edges – from 1 to 760 from all nodes (243 539). The explanation about the difference of value between γ_{out} and γ_{in} could be that the incoming links look less scalable.

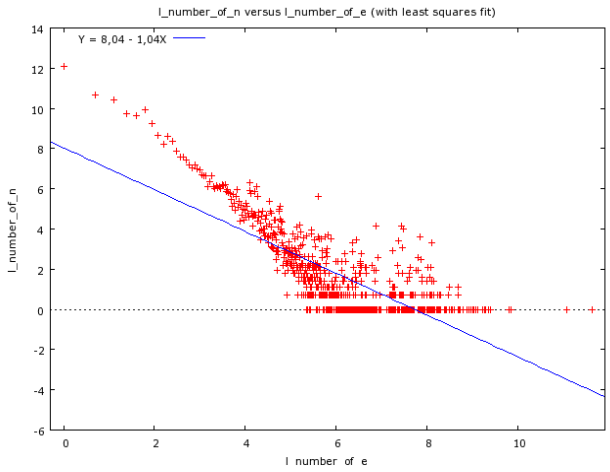


Figure 6 - Incoming $P_{in}(k)=cx^{-(1.04)}$

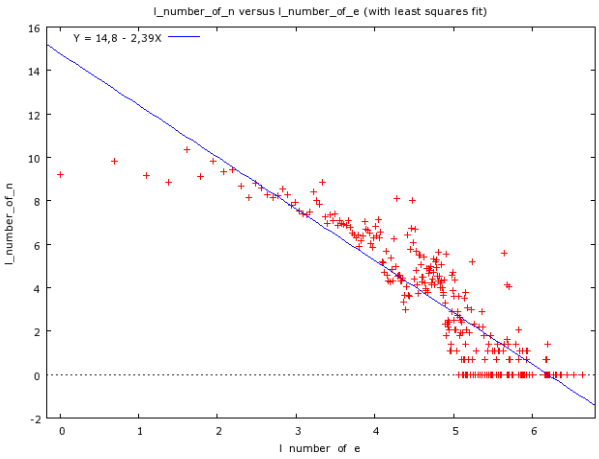


Figure 7 – Outgoing $P_{out}(k)=cx^{-(2.39)}$

3.2 Characteristics

Centers

Scale-free networks are specific by certain characteristics followed from power law. From this distribution we know that some nodes have high degree of interconnection. These centers carry the whole network in one piece *(*Figure 8*).*

Robustness and Failure

The robustness is given by number of alternative paths among majority of pair of nodes. The disintegration of system begins after eliminating 5 – 10 % of biggest nodes. In our case it means the elimination of the 18 745 biggest nodes from all nodes (374 928).

The simulations are showed in figures 9 and 10. Without centers the network becomes disintegrate, the islands are created (clusters without connection to others), the number of alternative paths is decreasing.

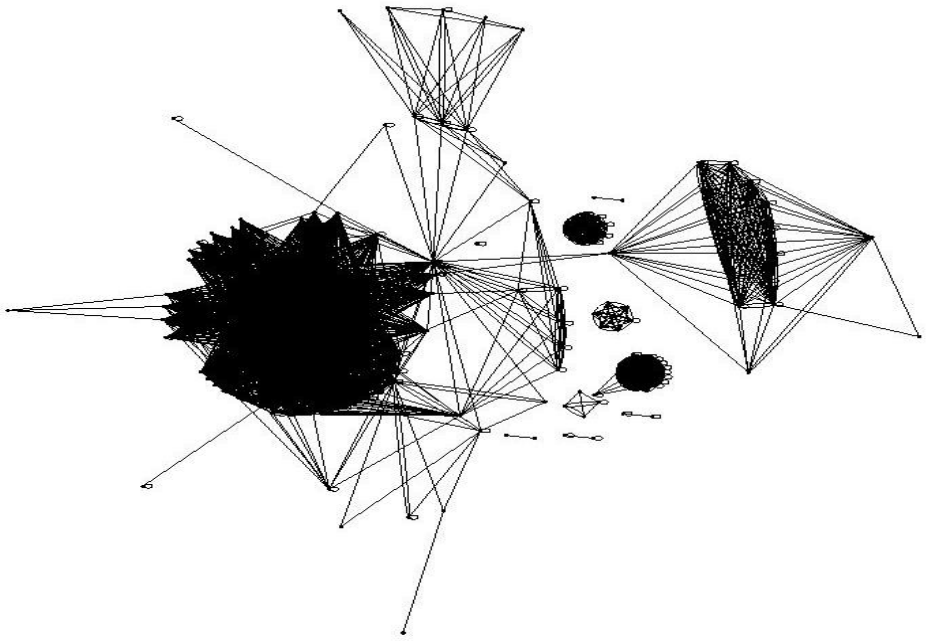


Figure 8 – Nodes above 3000 of incoming links. These 199 nodes (0.5% from all nodes) have 64% of all incoming links.

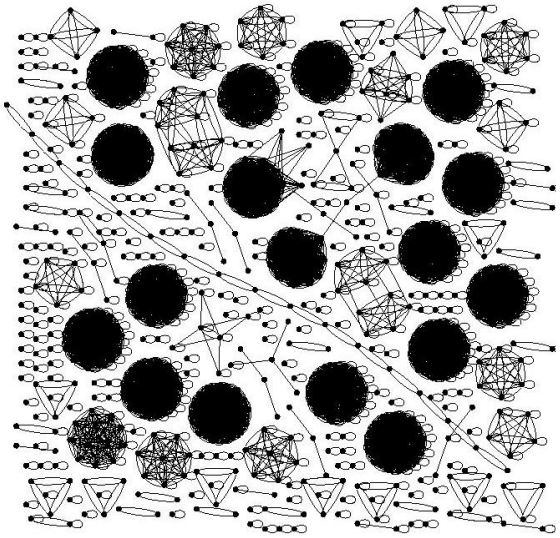


Figure 9 – Situation after eliminating 5% of the biggest centers – the interconnection of the following biggest nodes

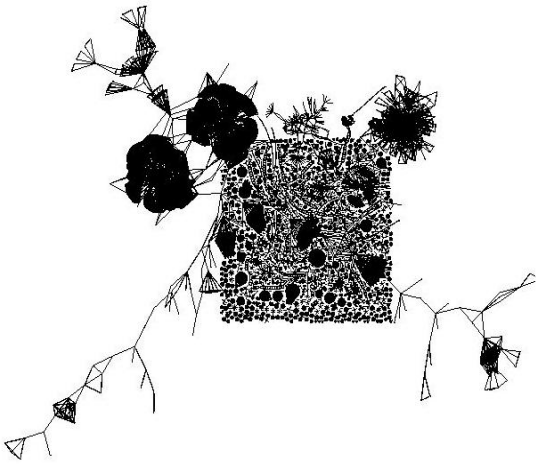


Figure 10 - Situation after eliminating 5% of the biggest centers – the interconnection of the nodes to the level 15 (edges)

REFERENCES

[1] Albert-László Barabási: *V pavučině sítí*, Paseka 2005, ISBN 80-7185-751-3

[2] Réka Albert, Hawoong Jeong, Albert-László Barabási: *Diameter of the World-Wide Web*, NATURE | VOL 401 | 9 SEPTEMBER 1999

[3] G. Bianconi, A.-L. Barabási: *Competition and multiscaling in evolving networks*, Europhys. Lett., 54 (4), pp. 436-442 (2001)

[4] A.L. Barabási, H. Jeong, Z. Néda, E. Ravasza, A. Schubert, T. Vicsek: *Evolution of the social network of scientific collaborations*, Physica A 311 (2002) 590 - 614

[5] A.-L. Barabási, Réka Albert, Hawoong Jeong: *Scale-free characteristics of random networks: the topology of the world-wide web*, Physica A 281 (2000) 69-77

[6] Lun Li, David Alderson, Reiko Tanaka, John C. Doyle, Walter Willinger: *Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications (Extended Version)*, Technical Report CIT-CDS-04-006, Engineering & Applied Sciences Division California Institute of Technology, Pasadena, CA, USA, Updated: October 2005

[7] David M. Pennock, Gary W. Flake, Steve Lawrence, Eric J. Glover, C. Lee Giles: *Winners don't take all: Characterizing the competition for links on the web*, PNAS 2002;99;5207-5211, February 2007

[8] http://www.nd.edu/~alb/