10th International Research/Expert Conference
"Trends in the Development of Machinery and Associated Technology"
TMT 2006, Barcelona – Lloret de Mar, Spain, 11-15 September, 2006

CLUSTER ANALYSIS OF A SCALE-FREE NETWORK

Nina Bijedic
University of Dzemal Bijedic, Faculty of Information Technologies
Mostar, Bosnia and Herzegovina

Senad A. Burak
University of Sarajevo, Faculty of Mechanical Engineering
Sarajevo, Bosnia and Herzegovina

ABSTRACT
In this paper we present an analysis of a cluster based inference in a particular computer network. The faculty forum on a real community server, where students and stuff share their knowledge and experiences, is used for this purpose. In order to better understand the structure of the network, we represent it as a graph, where vertices are represented by the members of the forum and the edges act as the links between the forum posts. As in many similar systems, this forum is organized in threads that are divided into sections (subjects), and sections are divided into groups (academic years). It is shown that the resulting network exhibits a scale-free distribution with large clustering coefficients following the small-world properties. As the clusters hold some important information about the nature of the network, we developed a special software agent that explores the background SQL database and automatically acquires the relevant information. Based on this data, detailed information including the graphs degree distribution, clustering coefficient, Laplacian, and normalized Laplacian eigenvectors and average distance are calculated. The resulting analysis gives us a better understanding of the nature of this particular network, which can be valuable information for the administrators.
Keywords: networking, programming, simulations

1. INTRODUCTION
An important part of the education process at the Faculty of Information Technologies in Mostar, Bosnia Herzegovina, is its system for distance learning. Among many modules and technologies used for this purposes, the faculty forum is one of the most popular, because it improves the communication between the students and staff and offers a valuable source of knowledge and experience to the users.
In order to better understand the knowledge sharing, we explore the topology of the forum's communication as a directed graph. In this model the users that posted their messages act as vertices and the edges of the graph are established if the user n_i posted an answer to a message posted by the user n_j. The resulting graph is then represented by an adjacency matrix. The following topology exploration includes the three operations: the distribution of the in-degree and out-degree vertex, the calculation of the average path length and the determination of the graph's clustering coefficient.

2. DATA ACQUISITION AND GRAPH MODELING
We explored the forum's data from the underlaying MS SQL Server 2000. During the observation period there were 602 registered users, but only 335 were active with the total number of 10,180 posts. In order to represent the graph by an adjacency matrix **A**, we developed a special software agent in C# language that aquired and automatically anlysed the data. The three most important fields

from the database were *UserID*, *PostID*, and *ParentID*. Using this information, we model the graph based on the following principle: every user who posted a post represents a vertex of the graph; **A**[*i*, j]=1 if user with *UserID*=*j* and *ParentID*(*j*)=*PostID*(*i*), and 0 otherwise. The resulting matrix was then used for further research.

3. KNOWLEDGE SHARING MODEL

3.1 Connectivity distribution

The next step was to investigate the distribution of the out-degree and in-degree of the graph. It turned out that the connectivity distribution for both in and out degree of this model followed the power-law, which means that the underlaying network fall in the scale-free category. Figure 1 illustrates this important observation.

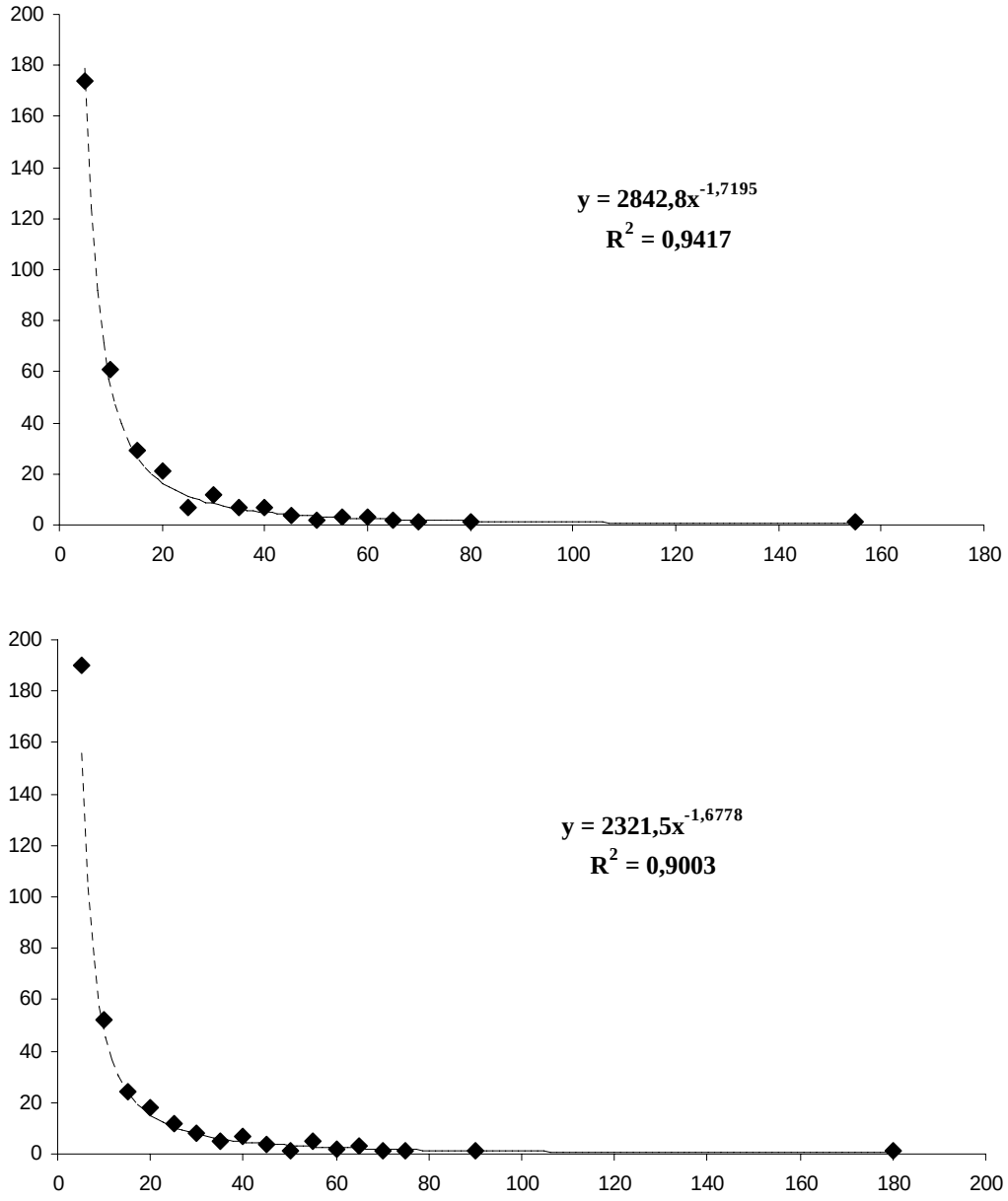

Figure 1. In-degree distribution (left) and out-degree distribution (right) with regression approximations. It is clear that both distributions follow power laws.

After basic statistical data analysis (**in-degree**: max=153, min=0, average=10.6149, stdev=15.529, median=5, mod=1; **out-degree**: max=180, min=0, average=10.6149, stdev=17.021, median=4,

mod=1) we decided to group the data and investigate further such new samples. It turned out that for both in and out degree the best regression approximation was the power regression with acceptable R^2. A better fitting (larger R^2) was obtained for denser data, representing in-degree distribution.

The obtained in-degree distribution follows the power law, given by

$$p_{in}(k) \propto k^{-\gamma}, \quad \gamma = 1.7195 \tag{1}$$

with the average in-degree coefficient

$$\langle k \rangle_{in} = 10.80645 \tag{2}$$

The regression approximation for out-degree distribution was

$$p_{in}(k) \propto k^{-\gamma}, \quad \gamma = 1.6778 \tag{3}$$

with the average out-degree coefficient

$$\langle k \rangle_{out} = 9.305556 \tag{4}$$

The results above indicated that we have a scale free model, but we also tried to make sure that this also implies a short average path length. Since it is obvious that the forum communication follows mechanism of constant grow and some sort of preferential attachment, we decided to explore theoretical aspects of our model as if it was BA model [3, 4].

3.2 Average path length

We tried to explore the average path length, both theoretically and using the *Dijkstra's* algorithm. However, it is shown that the resulting coefficient numbers differ significantly, which confirm the weakness of the theoretical approach. For theoretical approach we used formula suggested in [3] (with the author's remark that it gives better results for larger models). Our Dijkstra's average path length is given by

$$\underline{\langle l \rangle_{Dij} = 2.53} \tag{5}$$

and this result strongly implies that our model is a scale-free model. In the theoretical approach, as suggested in [4] for BA model, the resulting value for the average path length was $\langle l \rangle = 6.12$, which is larger more than twice compared to value in (5).

3.3 Clustering coefficient

In order to explore the clustering coefficient of our model, we used the theoretical approach suggested in [4] for BA model. Therefore, we used the following equation for our data

$$C(N) = \frac{m}{8} \cdot \frac{(\ln N)^2}{N} \tag{6}$$

After evaluating the above formula a rather moderate clustering coefficient is obtained, with the value of $c = 0.013$.

4. DISCUSSIONS

Having an opportunity to analyse a forum-based knowledge sharing at a real faculty distance learning web server we were able to study both, the preferential attachment and the constant growth

mechanisms and their impact to graph topology. The graph evolution observed after 110, 200, 300, 500, 1000, 2000, 3000, 5000, 7000 and 11080 posts indicated the scale-free characteristics of the model. The existence of hubs was clearly identified from the very beginning of the implementation of the web platform.

In this particular model, scale-free topology implies that most forum users are passive, being rather content with available information. This might imply that educators can use the forum as an additional tool for the direct type of communication with students. The short average path length (5), immanent to scale-free topology, which implies that we deal with a concentrated community inside the forum users. It also implies the existence of a small-world property of the model.

Theoretically obtained clustering coefficient only urged us to explore possibilities of cluster analysis for the model. Clusters in knowledge sharing model would emphasize the knowledge flow. In the further research we expect to discover clusters of students and educators grouped around the academic year subjects, as well around their professional interests. For our curriculum that would suggest both, the horizontal (by academic year) and vertical (professional interest) knowledge flow. We also expect to discover a cluster of so-called spammers that is those users whose posts are mostly not related to the usual topics of the forum.

To support the assumptions concerning the clusters, the eigenvalues and eigenvectors of adjacency and distance matrices are determined, where the distance matrix was obtained using the Dijkstra's algorithm. A lot of zeros and complex eigenvalues that are computed indicate that the network nodes can be indeed grouped together. This can suggest a possibility of finding the clusters of particular knowledge base.

5. CONCLUSION

Analysis of the topology of complex systems can give researchers a powerful tool in understanding the nature of many practical systems. In this case, the revealing scale-free topology of the active faculty forum means that we can better use it in the quality assessment and future development of the education systems. The potential treats to the functionality and security of the system can be predicted and numerical simulations can give a valuable insight in the system's behaviour.

6. REFERENCES

[1] A-L Barabasi and R. Albert 1999 Emergence of scaling in random networks, Science 286 509.

[2] Albert-Laszlo Barabasi, Reka Albert, Hawoong Jeong, Scale-free characteristics of random networks: the topology of the world wide web, Physica A 281 (2000) 69-77

[3] Agata Fronczak, Piotr Fronczak and Janusz A. Holyst, Average path length in uncorrelated random networks with hidden variables oai:arXiv.org:cond-mat/0407098 (2005-09-20) 2004.

[4] Konstantin Klemm and Vıctor M. Eguıluz, Growing scale-free networks with small-world behavior, Phys. Rev. E 65, 057102 (2002)