International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:9, 2014

Probabilistic Graphical Model for the Web

M. Nekri, A. Khelladi

Abstract**—**The world wide web network is a network with a complex topology, the main properties of which are the distribution of degrees in power law, A low clustering coefficient and a weak average distance. Modeling the web as a graph allows locating the information in little time and consequently offering a help in the construction of the research engine. Here, we present a model based on the already existing probabilistic graphs with all the aforesaid characteristics. This work will consist in studying the web in order to know its structuring thus it will enable us to modelize it more easily and propose a possible algorithm for its exploration.

Keywords**—**Clustering coefficient, preferential attachment, small world, Web community.

I. Introduction

THE Web, as growing complex system, has generated many problems which are mainly related to its organization, to the access to its content, Research Information, to the demanded information, etc.

Answering such questions requires the knowledge of the global structure of the Web i.e. all pages and hypertextual links composing the Web. However, all the studies that have been made in this field have not yet come to develop an approach to find the exact structure of the Web and this for various reasons in particular the one related to the dynamics of the web. So in the absence of real data, the probabilistic modeling, base on properties observed in the practice, is regarded as the best approach. It consists in making graph similar to the web graph, called the model graph, with the same properties and behavior as the real graph.

The model graph can be used to understand certain phenomena and describe the effective measurement procedures. Several research studies [1], [4]-[8] have been carried out on the web graph, these latter have made it possible to determine its statistical properties which are: The average distance between the web pages is short. If two pages point to a third one there will be a strong probability that the two pages are linked one to another. The last property is in relation either with the internal or with the external links of a web page. Indeed theses studies have shown that the minority of pages has a large number of links comparing to the other pages.

M. Nekri is with the Research Center on Scientific and Technical Information, 03 rue des frères Aissou Ben Aknoun Algiers, Algeria (phone: +213 661925673; fax: +21321912198; e-mail: nekri_mounira@ yahoo.fr).

A. Khelladi is with the University of Sciences and Technology Houari Boumediene USTHB, BP 32, 16111 El Alia, Bab-Ezzouar, Algiers, Algeria (e-mail: kader_khelladi@yahoo.fr).

II. Definitions

A. Definition of the Web Graph

The Web graph denoted $G = (V_w, E_w)$ is defined as follows:

$V_w = \{v_1, v_2, \ldots, v_n\}$ is the finite set of vertices. The v_i represent the Web pages and n the total number of the existing Web pages at a given time t;

$E_w = \{(v_i, v_j) / v_i \; and \; v_j \in V_w\}$ is the finite set of arcs. (v_i,v_j) represents the hypertextual links between the page v_i and the page v_j.

B. Definition of Random Graph

A random graph is generated by a random process. We start with an n isolated vertices and adding successive edges between them at random. The first model has been defined in 1959.

III. The Web Proprieties

A. Degree Distribution (Power Law)

In nature, we can see that major events are rare, while small events are many. For example, there are few big cities, but many small towns. These phenomena can be described by the power law which is a relationship between two sets x and y satisfying:

$$y = ax^k$$

Its probability distribution, known also as discrete distribution, is as follows:

$$P(x = k) = Ck^{-\gamma} \; with \; \gamma > 1.$$

Concerning the web, the power law can describe either the number of visits to a site or the distribution of degrees. Most studies have focused on the distribution of degrees and have shown that the probability that a vertex has a degree d, is a proportional to $\frac{1}{dx}$ with $x > 1$. This means that vertices with small degrees are much more than the vertices with high degrees. This result was confirmed by another study [4] on a 200 million pages providing other results relative to the in-degrees and the out-degrees distributions. For the in-degrees distribution, the probability that a vertex has an in-degree is proportional to $\frac{1}{dx}$ with $x = 2.1$ and $x = 2.7$ for the out degrees. Figs. 1-3[4] show:

- The in-degree distribution taken in May 99 and October 99 has a certain consistency with the power law which the exponent is equal to 2.09.
- The out-degree distribution taken in May 99 and October 99 which the distributions (those of May and October) depart somewhat from the law power (exponent = 2.72).

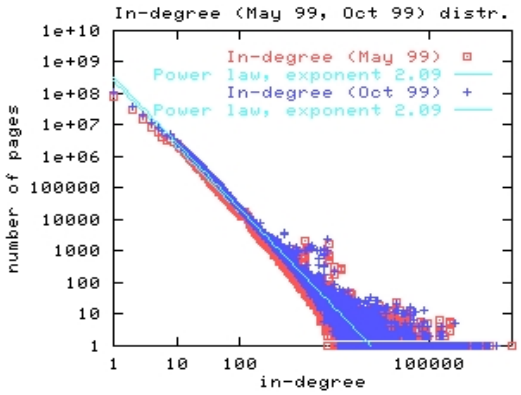

Fig. 1 In-degree distribution

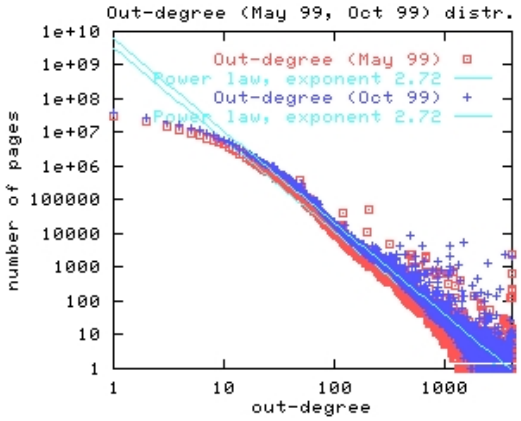

Fig. 2 Out-degree distribution

- The power law of the in-degree distribution compared to the Zipf law. Currently, researchers suggest that the pages with a low out-degree follow a different distribution: either a Poisson law or a combination between the Poisson law and the power law.

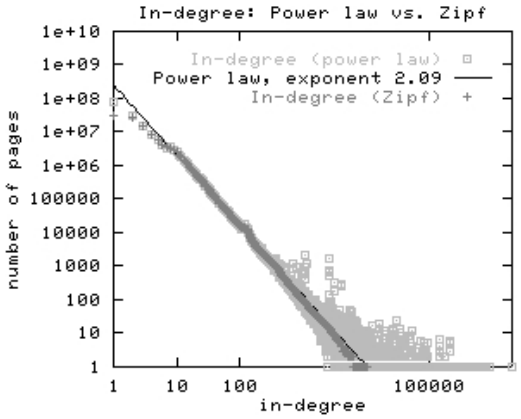

Fig. 3 The Zipf law

The diameter of a graph G is the longest, over all pairs (u, v), of the shortest path from u to v. Considering the graph G as the web graph, some research has determined that average distance is equal to 19. That is to say that we can reach from any web page, any other page using 19 links. Defined in [3], as the shortest path between two documents or as the smallest number of links to be taken to move from a document to another, the average distance of the Web graph was found equal to:

$$dist = 0{,}35 + 2{,}06\log(N)$$

The results showed that for a given N, dist follows the Gauss law, thus dist can be interpreted as the diameter of the Web.

For $N = 8.10^8$ the diameter of the web $< d >$ is equal to 18,59. This means that two documents which have been randomly taken on the Web are an average of 19 hits from one another. This small value of $< d >$, indicates that an intelligent agent, which can interpret the links and follow only those which are pertinent, can quickly find the desired information by surfing the Web, which is not the case for a robot that locates an information which is based on the "matching strings".

B. The Clustering Coefficient

The clustering coefficient is a measure defined in order to know if a graph is a small world. The clustering coefficient of a vertex in a graph quantifies how many this vertex and its neighbors are in a clique. More precisely, the clustering coefficient C_i for a vertex v_i is the proportion of links between its neighbors divided by the number of links that may exist between them. So, the clustering coefficient for a graph G is defined as follows:

$$C_i = \frac{|e_{jk}|}{k_i(k_i - 1)};\ v_j, v_k \in N_i,\ e_{jk} \in E$$

where k_i is the degree of vertex v_i and N_i is the set of its neighbors; This coefficient also represents the probability that two neighbors of the same vertex be themselves neighbors. The clustering coefficient of graph G denoted C is the average coefficient clustering of all its vertices.

$$C = \frac{1}{n}\sum C_i$$

C. The Small World

A graph G is said to be a small world if the most vertices are not neighbors but they can be reached by any vertex of G via a small number of edges.

Hence, a graph is considered as a small world if the average clustering coefficient C is significantly greater than the one in a random graph that is set up on the same set of vertices and if the graph has a low average distance.

IV. Some Models of Random Graphs Modeling the Web

After identifying the characteristics of the web, we studied different probabilistic graphs.

A. Erdős–Rényi Model

This is the first random model is defined in [5]as follows: let n be an integer number and p a real number such as $(0 \leq p \leq 1)$, the Erdős-Rényi model denoted $G(n,p)$ is a random graph with n vertices where each possible edge has a probability p of existing. This model allows the Poisson law degrees distribution, the average distance is low and it has almost no clustering coefficient.

B. The Barabási–Albert Model

Based on preferential attachment, the Barabasi-Albert model [2] is constructed as follows: from an existing vertex, the vertices are added one by one. The new vertices are connected to the vertices of the graph with probability P. This probability grows according to the degrees of vertices. This model is characterized by the power law degrees distribution, a low distance and a very clustering coefficient

C. Watts and Strogatz Model

Watts and Strogatz model [10] is a random that produces graphs with small-world properties, including short average path lengths and high clustering. Given the desired number of nodes n, the mean degree K (assumed to be an even integer), and a special parameter p, satisfying $0 \leq p \leq 1$ and $n \gg k \gg \ln(n) \gg 1$, the model constructs an undirected graph with n nodes and $\frac{nk}{2}$ edges in the following way:

1. Construct a regular ring lattice, a graph with n nodes each connected to K neighbors, $\frac{K}{2}$ on each side, such that, if the nodes are labeled $n_0 \ldots n_{n-1}$, there is an edge
2. (n_i, n_j) if and only if $|i - j| \equiv k(mod\ n)$ for some $|k| \in (1, \frac{k}{2})$ For every node $n_i = n_0 \ldots n_{n-1}$ take every edge (n_i, n_j) with $i \neq j$, and rewire it with probability p. Rewiring is done by replacing (n_i, n_j) with $(n_i, n_k)$where k is chosen with uniform probability from all possible values that avoid loops.

This process introduces $p\frac{NK}{2}$ links which can be connected to remote nodes of the original lattice. By varying p, we can interpolate between a lattice $(p = 0)$ and a random network $(p = 1)$ as shown in Fig. 4

(a) (b) (c)

Fig. 4 (a) Lattice network p = 0, (b) small-world network 0 < p < 1, (c) random network p = 1

Table I shows the values of the three properties of the above-mentioned models [3].

TABLE I
THE PROPRIETIES OF THREE MODELS

Models	Average distance	Power law distribution	Clustering coefficient
Erdos and Renyi	5.47	No	0.00002
Barabasi	*5.1*	Yes	0.0005
Watts and Stogatz	*11.23*	No	0.461

V. THE COMMUNITIES

The analysis of the large interaction networks shows that these networks share a common characteristic namely the existence of areas denser than others called ”communities” at a microscopic level. They are also sets of vertices which internal connections density is stronger than the density of connections to the outside. Researchers turn their attention to the communities because they are very useful for improving the performance of search engines. Meanwhile, the detection of communities can play an important role in the constitution of more complex algorithms. Indeed the large seize of the considered graphs is a major limit in terms of complexity of algorithms. In some cases using communities used to divide the graph can make cheaper separate calculations on each community [9].

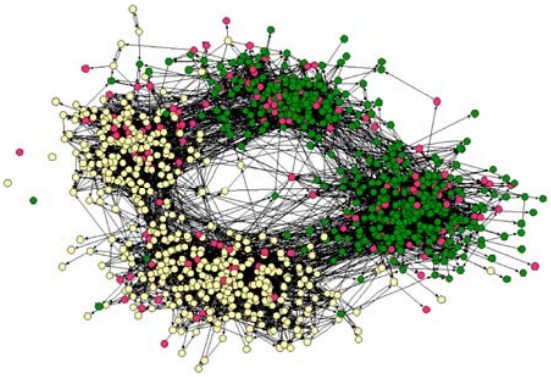

Fig. 5 Communities in any network

VI. THE PROPOSED MODEL

In order to modeling the web, our model is based on the follows models:

1. The model of small world in order to obtain fort clustering coefficient and low distance.
2. The Barabasi-Albert model for the power law.
3. The definition for communities in its HITS algorihm.

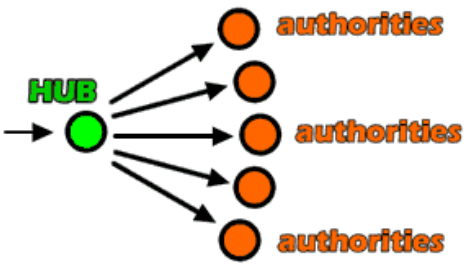

Fig. 6 Description of community

Hence, our model can be defined as follows:

First step:

With only one URL, we take a number n of URLs, then we do the following:

1. Make an exploration on each one of the URLs of the beginning.
2. Put the obtained communities in lattice.
3. Check if an authority existing in a community exists in another:

If this is the case then connect the two communities through a link; otherwise let it be.

4. If the graph obtained is a small world then end; else add or delete edges to obtain a small world.

Second step:"Procedure of adding vertices"

- Take the last vertex reached at the 1st exploration for each community.
- Make another exploration above each one of them, this will provide a significant number of vertices and therefore another partition communities, then return to the point "4" Or take a vertex and connect it to either an authority or a hit of any community. Our model will be represented as a lattice communities linked with a probability included between 0 and 1 in order to have a small world, and where each community will have one pivot and several authorities.

Remark

- If an authority has a neighbor v_i other than the pivot and the vertex v_i has several neighbors, then these neighbors will be put in a other community.
- If this neighbor v_i has less than 4 neighbors, then these neighbors will be taken in the first community.
- However, most of the time the authorities are the isolated vertices.

A. Validation of Model

To validate the obtained results, we have applied our model to the site of the Research Center on Scientific and Technical Information CERIST. We have obtained 2801 pages and 27020 links which were shared into 449 communities connected with probability equal to $\frac{1}{2}$. Each community has one pivot connected to several authorities. Concerning the clustering coefficient, it was found equal to 0.3978 and the average distance equal to 4.455. Compare these results with those of Table I, we can say that our model has the property of the power law; the distance and coefficient are close to theoretical values which are respectively equal to 7 and 0.446.

VII. Conclusion

In this work, we addressed the problem of the modeling of the Web as a directed graph. By Based on the three properties of the web, observed in practice, related to the average distance between the web pages, the number of links of a page, the clustering coefficient (a small world). We proposed a stochastic model which satisfies the abovementioned properties and can be used to determine the exact properties of Web in real time. To validate the obtained results, the proposed model was applied to the site of CERIST confirming that the model checked the three properties observed in practice. In addition, we intend to apply our model on a sample of considerable size (billion of vertices) and compare the empirical results with those of the existing models.

References

[1] P. Baldi, P.Frasconi and P.Smyth, "Modeling the Internet and the web, Probabilistic Methods and Algorithms," John Wiley and Sons, Ltd, Chichester, West Sussex, England, 2003.

[2] A.L. Barabasi, R. Albert, "Emergence of scaling in random networks," Science 286, 1999, pp.509–512.

[3] A.L. Barabasi, R. Albert, H. Jeong and G. Bianconi, "Power-law distribution of the World Wide Web," Science 287(2115), 2000.

[4] A. Broder, F. Kumer, S. Rajagopalan et all, "Graph structure in the Web," Computer Networks 33, 2000, pp.309–320.

[5] A. Bonato, "A survey of models of the web graph," Proceedings of Combinatorial and Algorithmic Aspects of Networking, 2004.

[6] J.L. Guillaume, M. Latapy, "Topologie d'Internet et de Web: mesure et modélisation," Actes du premier colloque Mesures de l'Internet, Nice, France, 2003.

[7] J.L. Guillaume, "Analyse statistique et modélisation des grands réseaux" PhD Thesis, Paris 7 university, France, 2004.

[8] M. Lyckova, I. Charon, L. et all. (2005), "Rapport sur le projet WEB-MOPT Optimisation et modélisation du graphe du Web," Département Informatique et réseaux. Groupe Mathématiques de l'Informatique et des réseaux.

[9] P.Pons, "Détection de communautés dans les grands graphes de terrain, "PhD Thesis, Paris 7 university, France, 2007.

[10] D. Watts and S. Strogatz "Collective Dynamics of Small-World Networks," Nature 393, 1998, pp 440-442.

M. Nekri obtained the engineer degree and the master degree in operation research from the University of Science and Technology USTHB Algiers, Algeria, respectively in 1996 and 2000. Currently, she prepares PhD Thesis in graph theory. Mounira Nekri is a permanent full-time researcher at the CERIST research center in Algiers, her researches focus on the following topics: graphs networks and dynamic graphs.

Mounira Nekri participated in many international conferences. She published more than 10 papers in international peer-reviewed journals and conference proceedings.