Analysis of the Greek Web-Space

T. Mchedlidze[1], A. Symvonis[1], M. Tzagarakis[2]

[1] National Technical University of Athens, School of Applied Mathematical & Physical Sciences, Athens, Greece.
symvonis@math.ntua.gr, mchet@central.ntua.gr
[2] Research Academic Computer Technology Institute, Patra, Greece.
tzagara@cti.gr

Abstract. In this paper, we report on our efforts to identify important Greek web sites on the web by analyzing the link structure of Greek web sites. Towards this end, five independent graph-theoretic methods have been deployed: hub and authority, page rank, Bow-Tie structure, core analysis and degree distribution. Our findings indicate that government, non-profit and educational sites are amongst the most important in the Greek web space.

1 Introduction

The World Wide Web contains enormous amount of information that is growing at a exponentially rate. Taking advantage of this move, diverse types of organizations with different social roles and missions put great efforts to relocate part of their activities into the world wide web. Hence, commercial sites attempt to market products and transact business on the web, news agencies and newspapers use it to inform the their readers, governments to deliver public services and educational organizations to provide virtual learning environments. Such transfer of activities from the real to the virtual world makes the World Wide Web effectively a mirror of today's society. This is in particular true if one limits the information available on world wide web to the level of a particular country.

Resembling a mirror of society, analysing the information available on the World Wide Web may give insights and hints about the zeitgeist of a country i.e. the intellectual and cultural climate of an era. The structural information in particular may give valuable information with respect to this. Identifying such trends may be beneficial for a number of reasons that include: archival reasons so that the cultural heritage of a country can be preserved for future generations and development reasons giving indications of communities that are either on the rise or on the fall. In this paper we report on our efforts to identify important Greek web sites on the web by analysing link structure of Greek web sites. Towards this, five independent graph-theoretic methods have been deployed: hub and authority, page rank, Bow-Tie structure, core analysis and degree distribution. Our findings indicate that government, non-profit and educational sites are amongst the most important in the Greek web space. The paper is organized as follows: Section 2 introduces the algorithms used to mine link information on the Greek web. In section three, we give an overview of the Greek web space and in section four we present the analysis and the results of our evaluation.

2 Methods Used in the Analysis

The analysis is based on five independent graph-theoretic methods: hub and authority, page rank, Bow-Tie structure, core analysis and degree distribution.

2.1 Hubs and Authorities

The *"Hubs and Authorities"* algorithm was introduced by Kleinberg [12] in an effort to identify in the world-wide web important pages that contain information related to a specific query topic s. In the literature, the algorithm is also referred to as the *HITS* algorithm (from the initials of the words in "Hypertext Induced Topic Selection").

For a specific query topic s, a web page is classified as an *authority on* s if it contains authoritative information about s, i.e., information that is important and truthful. For the same query topic s, a web page is classified as a *hub for* s if it links to web pages containing important information about s, i.e., a hub is one link away of important information on s. Having defined the notions of hubs and authorities, the HITS algorithm tries to estimate for every web page how good hub or authority it is. The algorithm is based on the simple idea that *"the page that points to many good authorities is a good hub and the page that is pointed to by many good hubs is a good authority"*.

The HITS algorithm assumes that a graph of the WWW has already been built. It issues a query about the search topic s on one of the text-based search engines and identifies in the graph the nodes corresponding to the top ranked web pages of the query result (typically the 100-200 top ranked pages are used; all these web pages contain the query text s in them). These nodes are referred to as the *root set* nodes. The root set of nodes is extended by including the nodes that link to its members (up to a predefined number) and the nodes that its members point to. The new set is referred to as the *base set*. The base set is expected to have some good properties: it is not very large, it is rich in relevant pages, it contains most of the strongest hubs and authorities. Thus, the HITS algorithm works on the subgraph G of the WWW graph induced by the nodes in the base set and tries to identify in it the hubs and the authorities.

For each node p of G, the algorithm computes two scores: a hub score $y^{<p>}$ and an authority score $x^{<p>}$. So, if $\boldsymbol{x}, \boldsymbol{y}$ are vectors that denote authority and hub values for all pages, then

$$x^{<p>} = \sum_{q:(q,p)} y^{<q>}$$

and

$$y^{<p>} = \sum_{q:(p,q)} x^{<q>}$$

or, in matrix form,

$$x \leftarrow A^T y \tag{1}$$

and

$$y \leftarrow Ax \tag{2}$$

The composition of (1) and (2) gives $x_k = A^T A x_{k-1}$ and $y_k = AA^T y_{k-1}$, where starting from some initial values x_0, y_0 the sequences $\{x_n\}_{n\in\mathbb{N}}$, $\{y_n\}_{n\in\mathbb{N}}$ converge to limits x^*, y^* respectively. It is a well known fact that x^*, y^* are the principal eigenvectors of $A^T A$ and AA^T, correspondingly. So, the principal eigenvectors of matrixes $A^T A$, AA^T reveal the importance of nodes and, in particular, x^* reveals how authoritative the web-pages are, while y^* reveals how good hubs they are. Kleinberg [12] points out that non-principal eigenvectors can be also used to partition the pages into groups of related hubs and authorities.

The method has been adopted in [8] to use the *href* text of every link, i.e., the text that appears "close to" the link in web page. The main idea is to assign to every link (p, q) a positive weight $w(p, q)$ that increases with the amount of topic-related text in the vicinity of the *href* from p to q.

An extension of HITS that incorporate both hyperlink and page content information was presented by Bharat and Henzinger [4]. They used weighted graph, where weights were computed from number of query term entries in the source page and the estimate of the document frequency of the query term in the World Wide Web. Also they tried to solve the problem of mutually reinforcing relationship between hosts using decreased edge weight if the page is linked with a lot of pages with the same domain name.

To improve the stability of HITS algorithm Ng, Zheng and Jordan [17] introduced two new versions of it. The first, "Randomized Hits" make usage of probabilistic parameter, which, in term of random walk on a graph, gives the probability of random jump of a surfer. This method is more stable to graph perturbations that the original HITS is. The second algorithm, "Subspace HITS", is based on the observation that subspaces spanned by a few eigenvectors may sometimes be stable even when individual eigenvectors are not. The method make use of multiple eigenvectors, presenting them as a single measure of authoritativeness.

Another improvement was made by Wang [19]. He proposed a new HITS matrix, elements of which are expressed by the probability of going from one vertex to another using a directed path if such exists.

Recently Awekar, Mitra and Kang [2] presented a new improvement of known HITS. They have changed the way the root set is built and, as a result, presented a selective expansion method which avoids topic drift and provides results consistent with only one interpretation of the query, even if the query is ambiguous. Also recently Maristella and Luka [16] presented a theoretical work on generalization of HITS algorithm.

HITS can be also used in creation of knowledge base of www [13]. The algorithm is applied on *extended bipartite cores* which can be thought to represent thematic fields. For a detailed description of the method see [13].

In our work we used the hubs and authority evaluation to extract the most dense community in the Greek web-space.

2.2 Page Rank

Page and Brin in [6] introduced *PageRank* which today lies at the heart of its software. PageRank is a method of ranking web-pages that gives an objective representation of the human notion of importance to every web-page. Independently of any search query, the web pages are ranked during an offline process. At the time the user makes a query, the relevant to this query pages are retrieved from the base and the best, accordingly to their pageranks, pages are presented to the user.

The intuition of the method is as follows: A web-page is supposed to be important if the pages that link to it are important as well. More formally: *The web-page has a high pagerank if the sum of pageranks of web-pages that link to it is high.* According to this definition, a web page has high pagerank when:

- It has a large number of incoming link from pages of low pagerank, or
- it has a small number of incoming links from web-pages of very high pagerank.

Let G denote the graph of the WWW pages and let $r(v)$ denote the pagerank value of web page v of G. If $In(v)$ is the set of pages that link to v and $deg(u)$ denotes the outdegree of a page u then, according to the above intuitive defenition, the pagerank of v may be computed as:

$$r(v) = \sum_{u \in In(v)} \frac{r(u)}{deg(u)} \tag{3}$$

Let T be the weighted adjacency matrix with weights defined by

$$\mathbf{T}_{i,j} = \begin{cases} \frac{1}{deg(i)} & \text{if i links to j;} \\ 0 & \text{otherwise;} \end{cases}$$

and $\boldsymbol{r} = (r(1) \dots r(N))$ be the vector of pagerank values for all the N pages of graph G. Then, (3) can be written as $\boldsymbol{r} = \boldsymbol{r}\mathbf{T}$. The equation is recursive and $\boldsymbol{r}$ can be computed starting from some initial value $\boldsymbol{r}_0$ and continue till convergence. To ensure the convergence of the method, table $\mathbf{T}$ has to be stochastic, i.e., the sum of the elements in every row has to be equal to 1, and the corresponding graph has to be strongly connected. So, a modified stochastic table $\tilde{\mathbf{T}}$, which is obtained from $\mathbf{T}$, is used instead. Table $\tilde{\mathbf{T}}$ is defined as follows. First define table $\mathbf{T}'$ to be

$$\mathbf{T'}_{i,j} = \begin{cases} \frac{1}{N} & \text{if } \forall j, \mathbf{T}_{i,j} = 0; \\ \mathbf{T}_{i,j} & \text{otherwise;} \end{cases}$$

Then, the *Google matrix* $\tilde{\mathbf{T}}$ is defined as $\tilde{\mathbf{T}} = c\mathbf{T}' + (1-c)\mathbf{E}$ where $\mathbf{E} = [1]_{n\times 1} \times \boldsymbol{\nu}^T$ and c is a positive constant smaller than 1. The pagerank $\boldsymbol{r}$ is then computed as a solution of the recursive equation $\boldsymbol{r} = \boldsymbol{r}\tilde{\mathbf{T}}$.

The Google matrix $\tilde{\mathbf{T}}$ can be easily explained in terms of a random walk on the web-graph. Imagine a surfer that walks on the vertices of the graph and let $T'_{i,j}$ represent the probability the surfer moves from vertex i to vertex j. If the surfer is at vertex i then, he either randomly uses one of the outgoing links of vertex i to move to a new vertex (with probability c) or, he moves to a random vertex of the web-graph (with probability $(1-c)$). Constant c is called the *damping factor*, and vector $\boldsymbol{\nu}$ is called the *personalization vector*. The personalization vector $\boldsymbol{\nu}$ is not necessarily uniform but can be biased towards some pages, so the surfer is more likely to go to some particular pages.

In contrast with a Hubs and Authorities, PageRank in its original form is not "topic-sensitive". A method to make PageRank sensitive was described by Haveliwala [10].

A very extended comparative work had been done by Langville and Meyer [14] where advantages and disadvantages of HITS and PageRank methods had been discussed.

The third eigenvector Web Information Retrieval method, after HITS and PageRank, called SALSA, was developed by Lempel and Moran in 2000 [15]. Like HITS algorithm the SALSA computes hubs and authorities scores, but make use of Markov Chains, like PageRank.

2.3 Bow-tie

While the HITS and the PageRank methods study the web-graph at a microscopic level, the *bow-tie* method examines it macroscopically. Broder and Kumar [7] observed that the web can be viewed as the bow-tie in Figure 1. According to them the web graph can be divided into several parts. The core of the web is a maximal *strongly connected component(SSC)* of it's graph. The left side of a bow-tie, named *IN*, represents the pages from which at least a path exists to some nodes in *SCC*. The right side of the bow-tie, named *OUT*, represents the pages which can be reached from nodes of *SCC*. The *TENDRILS* contains pages that are reachable from *IN*, or that can reach *OUT*, without passages through *SCC*. *IN* can be thought to consist of new pages that link to some "famous" ones but have not yet been discovered by *SCC*, while *OUT* can be thought to consist from some well known pages whose links point to internal pages only. As for *TENDRILS*, these pages are not yet discovered by the web, and they do not link to better-known regions.

2.4 Degree distribution

The macroscopic structure of the web-graph can be studied by examining the in-degree and out-degree distributions. For the WWW graph, Faloutsos et al. [9] discovered that the distribution of the in- and out-degrees of web-sites follows a power law. Namely, the number of pages with in-degree d is proportional to d^{k_I}, and the number of pages with out-degree d

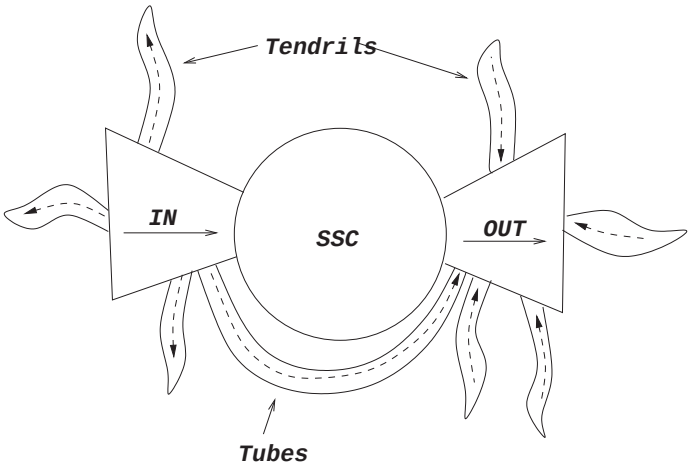

Fig. 1. Bow-Tie structure of the World Wide Web

is proportional to d^{k_O}, for some constants k_I and k_O. Kumar et al. [13] estimated k_I to be about -2.1 and k_O to be about -2.38. The growth of the exponent values can be explained as follows: *the higher the exponent value is, the smaller portion of the sites in the space monopolizes a large portion of the links included in space.*

In their effort to explain how the web grows, Kumar et al. [13] presented the (α, β)-model. In the (α, β)-model when a new page(vertex) is created, a new link is also inserted in the graph. α is a probability that the just-created edge has its target at the new vertex while, β is a probability that the just-created edge has its origin at the new vertex. Consequently, $(1-\beta)$ and $(1-\alpha)$ are the probabilities for newly created edge to have its origin and target at "old" pages. These two probabilistic parameters can be computed from the exponent values of the power law adapted to the web graph based on the equations:

$$k_I = -\frac{1}{1-\alpha} \qquad k_O = -\frac{1}{1-\beta}$$

Based on the estimates of Kumar et al. [13] for k_I and k_O, it follows that $\alpha = 0.52$ and $\beta = 0.58$.

In a related work, Ishimura et al. [11] studied the local web-space of several Japanese middle-size cities. They estimated values for the in-degree and out-degree exponents were $k_I = -1.14$ and $k_O = -1.39$, respectively. The corresponding probabilistic parameters were $\alpha = 0.12$ and $\beta = 0.28$. The small values of α and β (much smaller than 0.5), according to Ishimura et al. [11], reflected the fact that *a newly generated site is less likely to link to the existing sites in the local web-space the site belongs to.*

3 The graph of the Greek web space

3.1 Experimental setup

A crawler has been developed to retrieve pages from the Greek web. The crawler is written in C++ running on the linux operating system. A database maintains all pages retrieved from the world wide web. To speed up the gathering process the developed crawler supports multithreading and DNS caching reaching a throughput of about 6250 pages per hour.

We built the graph used in our analysis based on a crawl of the Greek web space. The crawl was constrained only to web pages in the ".gr" space. Dynamic as well as static pages were taken into consideration. In a breadth first search manner, the crawler discovered new pages and it explored their links. Pages outside the Greek web page were discovered but the crawler did not examined any of their links. As a result, our graph contained several nodes"at the boundary" of the Greek web space, all having out degree zero

Our intial set of URLs from which the crawler started harvesting pages consisted of 73400 distinct sites in the .gr domain. From these distinct sites the crawler downloaded a total of 1045563 web pages. We used the graphml format to represent it. Initially data were collected at a URL level, in particular all web-pages with different url have been saved for any site the crawler had found. For example, taking the case on Figure 2, for www.ypepth.gr we knew all pages with the same domain name and all outgoing links of these pages. In the PageRank and Hubs and Authorities evaluation, it made good sense not to take into account the secondary pages of web-sites[3]. So it was decided to restrict the data set to domain names.

Weighted Graph We call *internal links* those links which bind two pages with the same domain name, and *external links* those between two pages with different domain names.The first model we used had as vertices all different domain-names. Internal links were treated as being of low significance. Thus, we only modeled the external links between two domains, representing them by the weight of the directed edge connecting the two domains. For example, in Figure 2 the weight of the edge from www.ypepth.gr to www.e-yliko.gr is 2.

Note 1 *After restriction our graph had* 76506 *nodes and* 141791 *edges.*

4 Results of the evaluation

4.1 Important web-pages of our local web-space

Data Set One of our objectives was to apply the HITS and the PageRank algorithms to the Greek web-graph in order to identify the most important domains of our local web-space. We used the implementations in Visone[5] and Pajec[3].

[3] The links from secondary pages to primary ones, and viceversa, of the same domain do not generally reflect the public view, but just the view of web-page constructor.

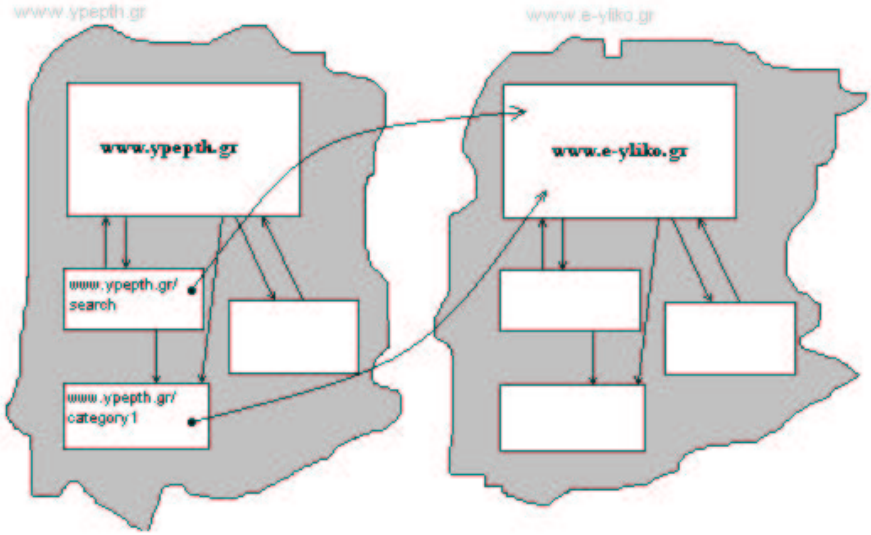

Fig. 2. Types of links.

We initially used Pajec to identify hubs and authorities. The choise of Pajec was due to the efficient implementation it provides, especially for large graphs. The three bottom lines of Figure 3, drawn by Pajek, displays the 12 best authorities and the 12 best hubs. The first bottom line contain the 8 best hubs, the second bottom line contains 4 domains that are both hubs and authorities, the third bottom line contains the remaining 8 best authorities, and the top line contains the rest of the domains.

To our surprise, 8 out of the 12 best authorities from are related to the newspaper "kathimerini". This fact does not agree with the intuitive notion of authoritativeness. The web-page "kathimerini" could be one of the best 12 authorities, but it shouldn't take 8 out of the 12 top positions. By carefully examining the data set we observed that the web-pages of "kathimerini" are placed at different domains. By restricting our initial data set to domain names, we expected that all internal links had been excluded, but "kathimerini", putting its pages on different domains, turned its (logical) internal links to appear as external. Table 1 contains all domains devoted to "kathimerini" and the numbers of links from one domain to another, or, in our implementation, the weight of the corresponding edge. By observing that more than 90% of the edges in our web graph have weights less than 10, the weights of the edges between "kathimerini" domains appeared to be very large. This explains why 8 of 12 best authorities are related to "kathimerini".

The "kathimerini" phenomenon made us suspect that our domain graph contains more domains corresponding to the same web site. We would like to eliminate these *duplications*, having each domain represented only once. To that end, we tried to locate domains that their structurally appear to be *almost identical* by focusing on the out-links of the domains. We treated two pages p and q as almost identical and we represented them as a single domain in our revised domain graph:

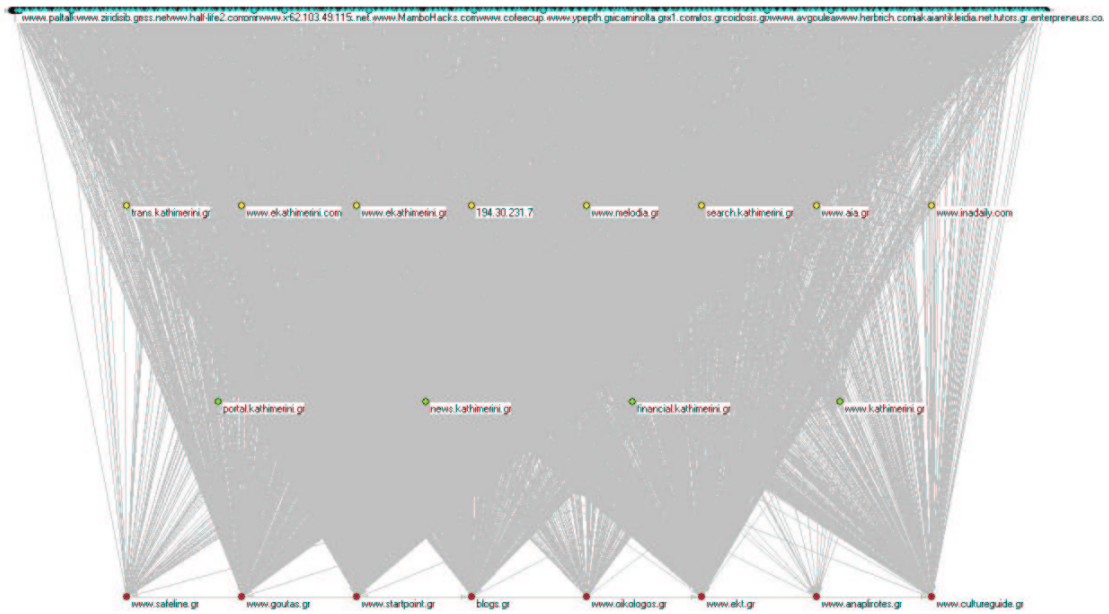

Fig. 3. Best 12 hubs and authorities as computed by Pajek for the initial weighted domain graph.

1. The number of outgoing links of p and q differ by 2: $|Out(p) - Out(q)| \leq 2$.
2. There are at most 6 different links out of the two domains (ignoring edge weights).

Note 2 *After filtering out almost-identical domains, our web-graph had* 76441 *nodes and* 139509 *edges.*

Below we present some indicative domains that were revealed to be almost-identical.

- The domains *radio.classicalmusic.gr* and *distribute.classicalmusic.gr* are different pages of one web-site which are different visually and have different theme, however, they have exactly the same outgoing links.
- The following are two images of the same page: *www.seareport.gr* & *www.beachreport.gr*, *www.xkatsikas.gr* & *www.alfavita.gr*, *www.winlife.gr* & *www.winforlife.gr*, *www.ianos.gr* & *www.books-online.gr*.

We found out that no two of the "kathimerini" domains had the same outgoing links. So, since we had no easy way to decide whether two domains are part of *the implementation* of the same web-site (and, thus, to solve the "kathimerini" and other unidentified similar problems), we decided to ignore the edge weigths and to experiment with a weightless graph.

Weightless Graph The Hubs and Authorities result for the weightless domain graph, as computed by Pajek, are presented in Tables 2 and 3.

From Domain	To Domain	Weight
www.kathimerini.gr	www.ekathimerini.com	744
www.kathimerini.gr	portal.kathimerini.gr	2266
www.kathimerini.gr	news.kathimerini.gr	37762
www.kathimerini.gr	www.ekathimerini.gr	7
www.kathimerini.gr	financial.kathimerini.gr	384
News.kathimerini.gr	www.kathimerini.gr	10102
News.kathimerini.gr	www.ekathimerini.com	2218
News.kathimerini.gr	portal.kathimerini.gr	9598
portal.kathimerini.gr	www.kathimerini.gr	3240
portal.kathimerini.gr	www.ekathimerini.com	339
portal.kathimerini.gr	news.kathimerini.gr	504
portal.kathimerini.gr	financial.kathimerini.gr	184
portal.kathimerini.gr	search.kathimerini.gr	75
portal.kathimerini.gr	photo.kathimerini.gr	122
financial.kathimerini.gr	www.kathimerini.gr	173
financial.kathimerini.gr	www.ekathimerini.com	17
financial.kathimerini.gr	portal.kathimerini.gr	142
financial.kathimerini.gr	search.kathimerini.gr	12
financial.kathimerini.gr	news.kathimerini.gr	12
financial.kathimerini.gr	trans.kathimerini.gr	72
News.kathimerini.gr	financial.kathimerini.gr	2460
News.kathimerini.gr	search.kathimerini.gr	1230
News.kathimerini.gr	trans.kathimerini.gr	7380
News.kathimerini.gr	photo.kathimerini.gr	1
www.kathimerini.gr	search.kathimerini.gr	186
www.kathimerini.gr	trans.kathimerini.gr	1116

Table 1. The links between the domains devoted to the "kathimerini" newspaper.

To get a better picture of the hubs and authorities results, we used the layered layout[4] provided by Visone [5]. The results of the visualization appear in Figures 4 and 5.

From the hubs-visualization (Figure 5) we observe that the difference between hub values is too large, while the authorities-visualization (Figure 4), reveals a large group of nodes displayed at the same layer. Trying to identify the reason for this fact, we noticed that there are 10114 nodes (out of the 76441 in the entire graph) that have identical authority value equal to 0,0081123. All these nodes have just one incoming link from *dir.forthnet.gr*. Knowing this, it is not surprising that the domain *dir.forthnet.gr* is the best hub.

The PageRank analysis of the weightless graph has been done using Visone[5]. The results are visualized in Figure 6 and the top 20 domains according to the PageRank evaluation are presented in Table 4.
Comparing the results of PageRank and HITS evaluation, we distinguish two types of popular domains:

[4] A layered layout emphasizes the importance of the nodes of a graph according to the weight assigned to the nodes. The layout places the nodes to different layers according to these weights. In our case we used the hub and authority weights.

Authority Weight	Web Page	Description
0,0164229	Europa.eu.int	European Union web Gate
0,014825	www.auth.gr	Aristotle University of Thessaloniki
0,0144888	www.parliament.gr	Hellenic Parliament
0,0144344	www.in.gr	Search Engine and Catalog
0,0144167	www.uoa.gr	National & Kapodistrian University of Athens
0,0142828	www.ntua.gr	National Technical University of Athens
0,0142745	www.aegean.gr	University of the Aegean
0,0142545	www.uoi.gr	University of Ioannina
0,013984	www.geocities.com	Yahoo Web Space
0,0138814	users.otenet.gr	Otenet
0,0137366	www.forthnet.gr	News,Catalog,Search Engine
0,0134783	www.mod.gr	Ministry of National Defence
0,0134387	www.europa.eu.int	European Union web Gate
0,0134349	www.pi-schools.gr	Pedagogic Institute& Min. of Education
0,0133442	www.teiath.gr	T.E.I. of Athens
0,0132942	www.asfa.gr	Athens School of Fine Arts
0,0131447	www.ekt.gr	National Documentation Center
0,012955	www.eie.gr	National Hellenic Research Foundation
0,0127159	www.grnet.gr	Greek Research &Technology Network
0,0125983	www.eap.gr	Hellenic Open University

Table 2. The 20 best authorities obtained from the analysis of the weightless domain graph.

Hub Weight	Web Page	Description
0,9923293	dir.forthnet.gr	Catalog
0,0334724	www.in.gr	Search Engine and Catalog
0,0272116	www.phantis.gr	Search Engine and Catalog
0,0249825	www.cso.auth.gr	Aristotle University of Thessaloniki Career Services Office
0,0244363	www.e-yliko.gr	Web Educational Portal,Min. of Education
0,024251	www.startpoint.gr	Entertainment Catalog
0,0237506	www.kosmikanea.gr	Career Catalog
0,0236865	www.e-yliko.sch.gr	Web Educational Portal,Min. of Education
0,0215675	www.ekt.gr	National Documentation Center
0,0214549	www.translatum.gr	Translations
0,0201813	www.e-businessforum.gr	Business Forum of Min. of Development
0,0200832	Users.lar.sch.gr	Hellenic School Web
0,0192014	www.kalamaria.gr	Municipality of Kalamaria
0,0170332	www.de.sch.gr	Hellenic Scholar Network,Min. of Education
0,0164346	www.noc.ntua.gr	Network Center of NTUA
0,0162967	www.plefsis.gr	Educational Portal
0,0148821	www.xkatsikas.gr	Educational Network
0,0147811	www.uoi.gr	University of Ioannina
0,0146873	www.nad.gr	Municipality of Dodekanisa
0,0141403	www.robby.gr	Search Engine and Catalog

Table 3. The 20 best hubs obtained from the analysis of the weightless domain graph.

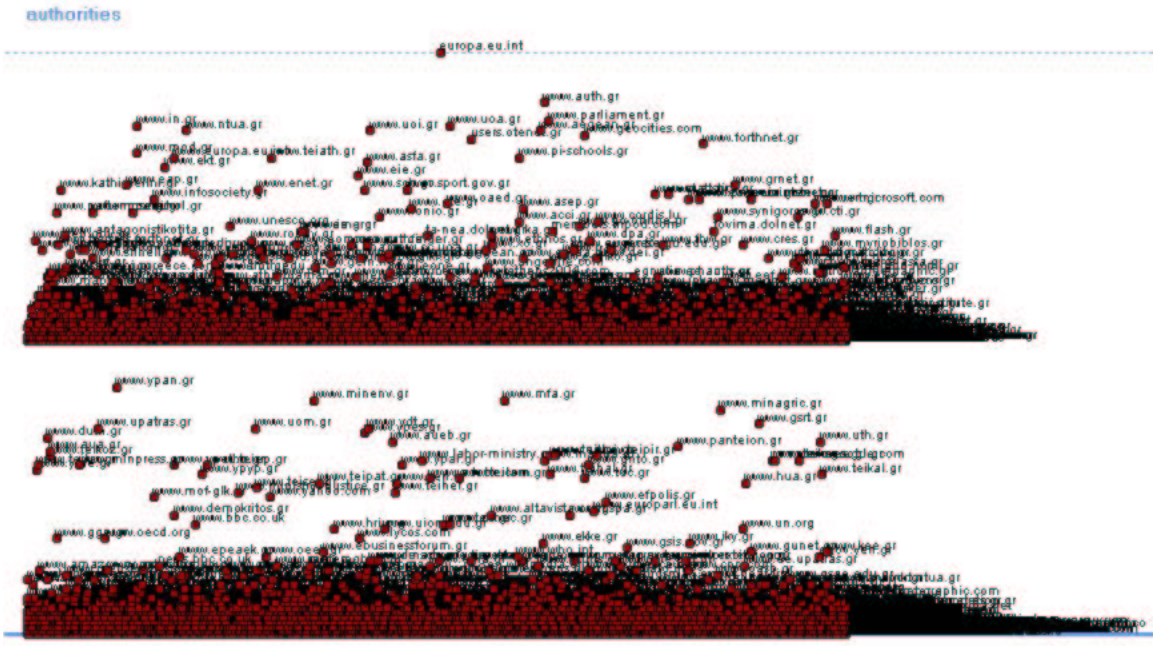

Fig. 4. Authority Weights. Layered Layout.

– Domains that are very popular (high pagerak) and are authorities for their contained information; these domains are listed in Table 5. [5]
– Domains that are very popular portals and search engines; these domains are listed in Table 6. [6]

The results of Table 5 and Table 6 are partially consistent with top 100 pages list given by ALEXA [1]. At ALEXA's web-page the top 100 sites of the Greek web space are ranked according to traffic data of Alexa Toolbar users. The domains that are highly ranked both by our experiment and Alexa are : *www.uoa.gr, www.ntua.gr, www.forthnet.gr* and *www.in.gr*.

4.2 Macroscopic Structure

Bow-Tie Analysis The bow-tie analysis of the weightless graph was facilitated by Pajek. By an analysis similar to that by Broder and Kumar [7], we found out that the *SCC(strongly connected component)* and the *IN* components of our graph were rather small. Table 7 presents the sizes of all components.

[5] The order of every site in this list is based on the sum of its authority and pagerank order numbers.
[6] The order of every site in this list is based on the sum of its authority and hub order numbers.

Fig. 5. Hub Weights. Layered Layout.

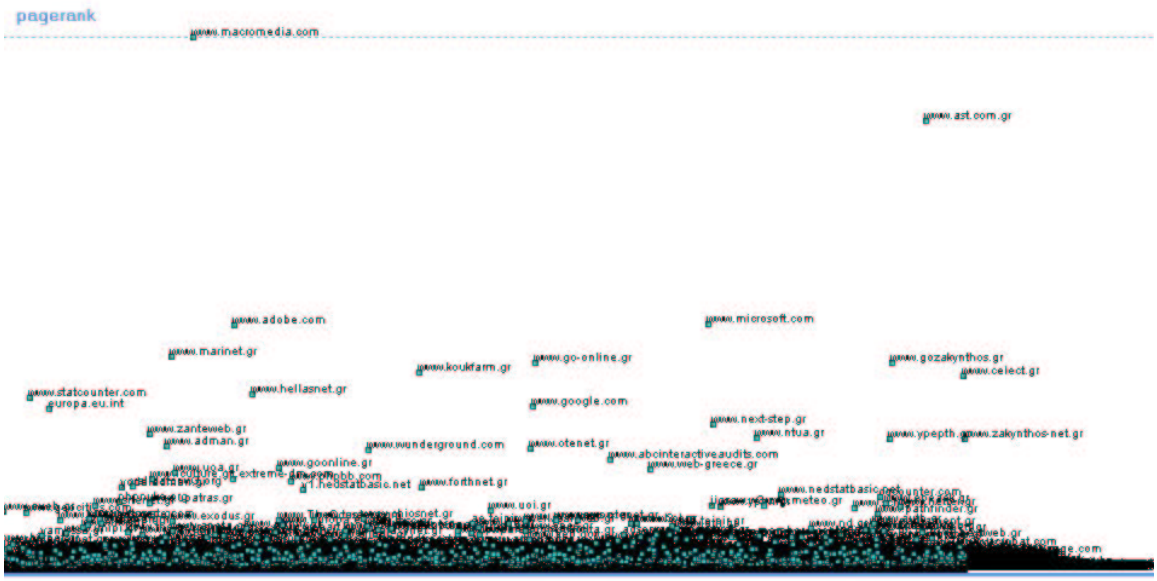

Fig. 6. Pagerank. Layered Layout.

Order	URL	Description
1	www.macromedia.com	Adobe
2	www.ast.com.gr	AST
3	www.microsoft.com	Microsoft
4	www.adobe.com	Adobe
5	www.marinet.gr	Portal
6	www.gozakynthos.gr	Zakynthos web-page
7	www.go-online.gr	Go-online(Diktiothite)
8	www.celect.gr	Celect
9	www.statcounter.com	Web page promotion company
10	www.hellasnet.gr	ForthNET-products
11	www.google.com	Google
12	www.europa.eu.int	European Union web Gate
13	www.next-step.gr	Development and Organization of Comp.
14	www.zanteweb.gr	Zakynthos web-page
15	www.ntua.gr	NTUA
16	www.zakynthos-net.gr	Zakynthos web-page
17	www.ypepth.gr	Min. of Education
18	www.adman.gr	Adman-promotion in internet
19	www.wunderground.com	Weather Portal
20	www.otenet.gr	OTENET
21	www.goonline.gr	Go-online(Diktiothite)
22	www.web-greece.gr	e-traveller
23	www.uoa.gr	National and Kapodistrian University of Athens
24	www.culture.gr	Hellenic Culture
25	www.forthnet.gr	ForthNET

Table 4. The 25 top domains according to pagerank values obtained from the analysis of the weightless domain graph.

www.europa.eu.int
www.ntua.gr
www.uoa.gr
www.forthnet.gr

Table 5. Most popular pages of Greek web.

www.in.gr
www.ekt.gr
www.uoi.gr

Table 6. Most popular portals and search engines of Greek web.

Core The subgraph G' of graph G is called a k-core, if every node of G' is connected with at least k nodes in G' and G' is a maximal. A core of the domain graph is thought to identify domains with strong ties of some form. So, we tried to identify the "heart" of the Greek web space, as it is represented by a core.

Applying the core search algorithm, implemented by Batagelj[3], and focussing on the in-degree, we found that the maximal core of the studied graph is 11-core with 66 nodes. It is evident that the domains in the core are related to Greek government, non-profit and educational institutions. The domains of the core are presented in Table 8.

Component	Vertex Number
SCC	2898
IN	15
OUT	73403
TUBES	0
TENDRILS	72
OTHERS	53

Table 7. Bow-Tie Analysis. Components Sizes.

URL	URL	URL
www.ypepth.gr	www.europa.eu.int	www.grnet.gr
www.gsrt.gr	www.gunet.gr	www.auth.gr
www.minagric.gr	www.primeminister.gr	www.culture.gr
www.sport.gov.gr	www.yen.gr	www.ministryofjustice.gr
www.ypyp.gr	www.minpress.gr	www.labor-ministry.gr
www.ydt.gr	www.mof-glk.gr	www.minenv.gr
www.ypes.gr	www.mathra.gr	www.mfa.gr
www.mod.gr	www.ypan.gr	www.ypai.gr
www.yme.gr	www.teilar.gr	www.teimes.gr
www.teipat.gr	www.teipir.gr	www.teiser.gr
www.teihal.gr	www.parliament.gr	www.et.gr
europa.eu.int	www.adobe.com	www.aegean.gr
www.cordis.lu	www.asfa.gr	www.teithe.gr
www.teikoz.gr	www.teilam.gr	www.teiath.gr
www.teiep.gr	www.teikal.gr	www.uoa.gr
www.un.org	www.europarl.eu.int	www.oecd.org
www.infosociety.gr	www.gnto.gr	www.uoi.gr
www.uom.gr	www.ntua.gr	www.aueb.gr
www.unipi.gr	www.upatras.gr	www.government.gr
www.efpolis.gr	www.neagenia.gr	www.ypetho.gr
www.uth.gr	www.aua.gr	www.teikav.edu.gr
www.panteion.gr	www.hua.gr	www.duth.gr

Table 8. The 66 domains of the 11-Core (based on the in-degree of the nodes).

Power-Law As mentioned in Section 2, the in- and out-degrees of vertices of the World Wide Web graph are power-law distributed. This fact was proved also for the local-web space of Japanese cities. In this section, we describe the results we have obtained by a similar analysis. For every number $i \in \mathbb{N}$ (let us call parameter i the *"in-degree"*) we computed the number of nodes (we call this parameter *"number of domains"*) with in-degree i that are contained in our graph. The do the same for the out-degrees.

Adapting the linear model to ln(in-degree), ln(number of domains), where the first is an independent and the second is a dependent parameter, we 've got that the best adapted lines were (see Figures 7 and 8):
For in-degree: $\ln(\text{number of domains}) = 4,00 - 1,91 \ln(\text{in-degree})$, while

for out-degree: $\ln(\text{number of domains}) = 2,34 - 0,903 \ln(\text{out-degree})$.
We note that the power law explains the in-degree distribution by 90%, while the out-degree distribution is explained by 72%. The 72% of the out-degree may be explained by the very big percentage of pages with out-degree 1. Thus, the exponent values of power-law for our graph are $k_I = -1.91$ and $k_O = -0.9$.

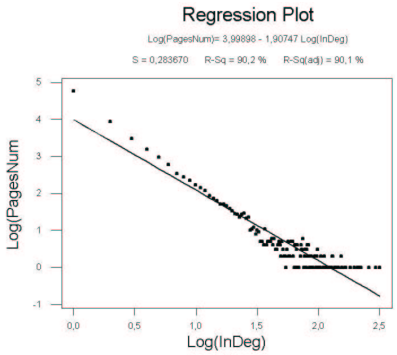

Fig. 7. Power-Law for in-degree.

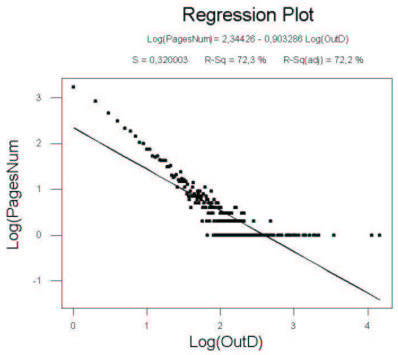

Fig. 8. Power-Law for out-degree.

Recall that the exponent values of Kumar et al. [13] for the WWW are: $k_{I_{WWW}} = -2.1$, $k_{O_{WWW}} = -2.38$, while the values of Ishimura et al. [11] for local web-space Japanese cities are $k_{I_{Japan}} = -1.14$ and $k_{O_{Japan}} = -1.39$. Refraining from drawing any conclusions for the out-degrees due to the low-percentage of data explanation by the model, we conclude that *in the Greece web space there are a lot of domains that are pointed to by a large number of other domains*. This conclusion can be made by contrasting our $k_I = -1.91$ with $k_{I_{Japan}} = -1.14$ and noting that their difference is greater than the difference of $k_I = -1.91$ and $k_{I_{WWW}} = -2.1$. In simple words, we can conclude that the growth of the Greek web-space is more likely to resemble the growth of global web than the growth of web of one town.

Interpreting the exponent value of in-degree in terms of (α, β)-model, we get that $\alpha = 0,48$, therefore a new created link has approximately the same probabilities to show a new web-page of our local space or to show at already existing web-page.

5 Conclusion

We have analyzed the graph of the Greek web space by applying social network techniques. Our future work includes studying the dynamics of the Greek web, focussing on the evolution of the Greek domain graphs resulting from successive crawls.

References

1. http://www.alexa.com/site/ds/
2. A. C. Awekar, P. Mitra, J. Kang. Selective Hypertext Indused Topic Search. *15th International World Wide Web Conference.* Edinburg, Scotland, May, 2006.
3. V. Batagelj, A. Mrvar, M. Mutzel. Graph Drawing Software. *Pages 77-103, Springer.* Berlin 2003.
4. K. Bharat, M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *In Proceeding of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM Press, New York, NY, 104-111, 1998.
5. U. Brandes, D. Wagner. VisOne-Software for Visual Social Network Analysis. *2002-first version of visone. Web Site: http://visone.info/*
6. S. Brin, L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems, 30, 1-7, 107-117.* 1998.
7. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Weiner Graph Structure in the web. *9th International World Wide Web Conference.* 2000.
8. S. Chakrabarti, B. Dom, P. Raghavan, S. Rajogopalan, D. Gibson, J. Kleinberg. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *In Proceeding of the 7th World-Wide Web conference, Amsterdam.* 1998
9. M. Faloutsos, P.Faloutsos, C. Faloutsos On Power-Law Relationship of the Internet Topology *Proceedings of SIGCOMM'99, pp. 251-261.* August 1999.
10. T. Haveliwala Topic-Sensetive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. *In IEEE Transactions on Knowledge and Data Engineering.* 2003.
11. Y. Ishimura, K. Shin, N. Kamibayashi. Graph-theoretic nature of local web-spaces. *Proceeding of the 16th International Workshop on database and Expert Systems Applications* 2005.
12. Jon M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM (JACM), v.46 n.5, p.604-632.* September 1999.
13. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins. Extracting Large Scale Knowledge bases from the Web. *Proceeding of the 25th VLDB Conference, Edinburgh, Scotland.* 1999.
14. A. Langville, C.D. Meyer. A Survay of Eigenvector Methods for Web Information Retrieval. *SIAM Rev.47, 1, 135-161.* 2005.
15. R. Lempel, S. Moran. The Stochastic Approach for link-structure analysis(SALSA) and the TCK effect. *In The Ninth International WWW Conference.* May 2000.
16. A. Maristella, P. Luca. A Theoretical Study of a Generalized Version of Kleinberg's HITS Algorithm. *Information Retrieval, Volume 8, Number 2, pp. 219-243.* Springer, April 2005.
17. A. Ng, A. Zheng, M. Jordan. Stable Algorithms for Link Analysis. *In Proceeding of the 24th Annual International ACM SIGIR Congerence on Research and Development in Information Retrieval.* ACM Press, New York, NY, 258-266,2001.
18. Page, Brin, Motwani, Winograd. The PageRank Citation Ranking: Bringing Order to the Web. *Proceeding of AS/S'98, Stanford University.* 1998.
19. M. Wang. A Significant Improvement to Clever Algorithm in Hyperlinked Environment. *In The Eleventh International WWW Conference.* May, 2002.