

# Structural Analysis of the Web

Pratyus Patnaik<sup>1</sup>, Sudip Sanyal<sup>1</sup>

<sup>1</sup> Indian Institute of Information Technology, Allahabad, India  
pratyus@ug.iiita.ac.in, ssanyal@iiita.ac.in

**Abstract.** World Wide Web has evolved exponentially since its inception. Today, it has become important for the algorithms of the web applications like searching, web-crawling, community discovery to exploit the information hidden in the hyperlink graph of the Web. This is the main driving force of the webmining community. We present in this paper an extensive analysis of the web, purely, based on the graph analysis algorithms. Prior works in the graph based analysis of the Web have been based on certain criteria and themes. But, we believe, for the web, which has evolved stochastically, only a pure graph based analysis can give us the true insights. We have carried out our analysis on isomorphic subgraphs of the Web to arrive at our conclusions. We reaffirm that the Web, is indeed a Fractal. Each structurally isomorphic subgraph shows the same characteristics as the Web and follows the classical Bow-tie model.

**Keywords:** Web Mining, Subgraph Isomorphism, Graphs, Fractals.

## 1 Introduction

The Web is an ever growing repository of large amount of information, spread across several servers in a complicated network. With the advent of Peer2Peer, web publishing every user is a disseminator of information. With every new webpage, a new node and with every link, a new directed edge is added to the web graph.

This growth in the size of the web presents an important task of mining and extracting relevant information from the hyperlink graph which can be exploited by simple searching and crawling algorithms as well as by advanced web applications such as web-scale data mining, community extraction, construction of indices, taxonomies, and vertical portals. In the recent past many application have surfaced which exploit the knowledge of the hyperlink structure of the web. Few such applications are the advanced search applications [4, 5, 6], browsing and information foraging [7, 8], community extraction [9], taxonomy construction [10].

Substantial amount of work has been carried out in the recent past [1, 2, 3]. [1] is very theoretical, and proposes stochastic models to explain the hyperlink structure of the Web. [3] talks about Small World Network and Scale Invariance in the structure of the Webgraph. It also proposes the classical bow-tie structure based on the various graph parameters (discussed in the next section). In [2], Kumar et. al. have extended the above model and have shown the self-similarity in the Web i.e., each thematically unified region displays the same characteristics as the Web at large. They have

characterized sub graphs as collections of Web pages that share a common attribute like keyword, content, location and some randomly generated subgraphs.

But, we believe to capture the true insights on the structure of the web, which has evolved stochastically over the period of its existence, we need to make use of pure graph based sub graph isomorphism algorithms. We applied iterative subgraph isomorphism algorithm on the webgraph to get the subgraphs. We then calculated various graph analysis parameter for those subgraphs. We found that each structurally similar subregion shows the same characteristic as the web and this holds for a number of parameters.

In the subsequent sections we have described our experiments, our results and finally the conclusions. But before delving deeper into the experiment, in the next section we briefly discuss the hyperlink webgraph, the graph parameters and the subgraph isomorphism algorithm.

## **2 Terminologies and Algorithm**

### **2.1 Web Graph**

Our view of the Web as a graph is same as [2] i.e. we ignore the text and other content in pages, focusing instead on the links between pages. In the terminology of graph theory [11], we refer to pages as nodes, and to links as edges. In this framework, the Web is a large graph containing over a billion nodes, and a few billion edges.

### **2.2 Graph Terminologies**

A directed graph consists of a set of nodes, denoted as  $V$  and a set of edges, denoted as  $E$ . Each edge is an ordered pair of nodes  $(u, v)$  representing a directed connection from  $u$  to  $v$ . The outdegree of a node  $u$  is the number of distinct edges  $(u, v_1), \dots, (u, v_n)$  (i.e., the number of links from  $u$ ), and the indegree is the number of distinct edges  $(v_1, u), \dots, (v_n, u)$  (i.e. the number of links to  $u$ ). A path from node  $u$  to node  $v$  is a sequence of edges  $(u, u_1), (u_1, u_2), \dots, (u_n, v)$ . As the graph is directed, a path from  $u$  to  $v$  does not imply vice-versa. The distance from  $u$  to  $v$  is  $n+1$ , for the smallest value of  $n$ . If no path exists, the distance from  $u$  to  $v$  is infinity. If  $(u, v)$  is an edge, then the distance from  $u$  to  $v$  is 1.

### **2.3 Graph Analysis Parameters**

A brief description of the parameters we have used in the analysis of the Web graph:

**Characteristic Path Length and Diameter.** The characteristic path length defines the typical distance from every node to every other node. The diameter represents the maximum possible distance between all the pair of reachable nodes. The Characteristic path length is calculated by finding the median of the means of the shortest paths from each node to every other node.

**Clustering Coefficient.** It is defined as the mean of the clustering indices of all the nodes in the graph. To find it, we find the neighbors of the node and then find the number of existing links amongst them. The ratio of the number of existing links to the number of possible links gives the clustering index of the node.

**Centrality and Centralization.** The degree centrality for a node is defined as:

$$C'_D(p_k) = \frac{\sum_{i=1}^n (a(p_i, p_k))}{n - 1}$$

where  $a(p_i, p_k)$  is 1 iff  $p_i$  and  $p_k$  are directly connected in the direction from  $p_i$  to  $p_k$ . The degree centrality of a point is useful as an index of a potential communication ability.

**Degree Centralization.** The centralization of a network is calculated as the ratio of the centrality of each node of the network with a star network of the same size.

**Betweenness Centrality.** It is based upon the frequency with which a point falls between pairs of other points on the shortest or geodesic paths connecting them.

**Closeness Centrality.** It is related to the control of communication in a somewhat different manner. A point is viewed as central to the extent that it can avoid the control potential of others.

## 2.4 Web Graph Characteristics

As mentioned earlier Small World Network and Scale Invariance are two important characteristics reported in earlier works [1, 2, 3].

**Small World Network.** It is a complex network in which the distribution of connectivity is not confined to a certain scale, and where every node can be reached from every other by a small number of hops or steps. The Web was shown to exhibit this characteristic first by [3], since then many have reinforced this assertion.

**Scale Free Networks.** Scale-free Networks, are the outcome of random construction processes. One of their common property is that the vertex connectivities follow a scale free power-law distribution. Power-law distribution states that for a positive integer, the probability of the value  $i$  is directly proportional to  $i^{-k}$  for a small positive number  $k$ . Scale free Networks are generic and are preserved under random degree preserving rewiring. They are Self Similar and Domain Independent [12].

Scale-free networks usually contain centrally located and interconnected high degree nodes, which influence the way the network operates. For example, random node failures have very little effect on a scale-free network's connectivity or effectiveness; but deliberate attacks on such a node can lead to a complete break down [12].

## 2.5 Subgraph isomorphism Algorithm

Although, graph isomorphism is a classical problem, this NP hard problem has no foolproof algorithm yet. Algorithms that we considered for our analysis were FSG [13], gFSG [14], gSPAN [15], GREW [16] and SUBDUE [17]. All these graph algorithms, except the last one, cannot handle graphs of more than 1000 nodes.

SUBDUE is heuristics based and hence is able to work on large unlabeled graphs. But, it gives an approximate result. Input to the SUBDUE system was the single web graph. SUBDUE outputs substructures that best compress the input dataset according to the Minimum Description Length (MDL) [18] principle. MDL has the fundamental idea that any regularity in a given set of data can be used to compress the data, i.e. to describe it using fewer symbols than needed to describe the data literally [17]. Since we wanted to select the hypothesis that captures the most regularity in the data, we looked for the hypothesis with which the best compression can be achieved.

SUBDUE performs a computationally-constrained beam search which begins from substructures consisting of all vertices with unique labels. The substructures are extended to generate candidate substructures. Candidate substructures are then evaluated according to how well they compress the Description Length (DL) of the dataset. Compression takes place by replacing all the subgraph instances by a single vertex. The DL of the input dataset  $G$  using substructure  $S$  can be calculated using the following formula,

$$I(S) + I(G|S)$$

where,  $S$  is the substructure used to compress the dataset  $G$ .  $I(S)$  and  $I(G|S)$  represent the number of bits required to encode  $S$  and dataset  $G$  after  $S$  compresses  $G$ . This procedure repeats until all substructures are considered or user-imposed computational constraints are exceeded. At the end of the procedure SUBDUE reports the best compressing substructures. This can also be carried out iteratively.

### 3 Experimental Details

We crawled webdata using WebMine [19]. We collected a hyperlink graph of websites of 1 million nodes and nearly 2 million edges. We have restricted our analysis to website graph because of the limitation of the SUBDUE algorithm in handling very large graphs. The graph was partitioned to make sample graphs of size ranging from 30K nodes to 100K nodes. Then SUBDUE algorithm was used to generate subgraphs of the sample graphs and also the complete webgraph. SUBDUE iteratively ran thrice for each sample to generate subgraphs at three levels of compression. WebMine tool was again used to compute various graph analysis parameters for the web graph.

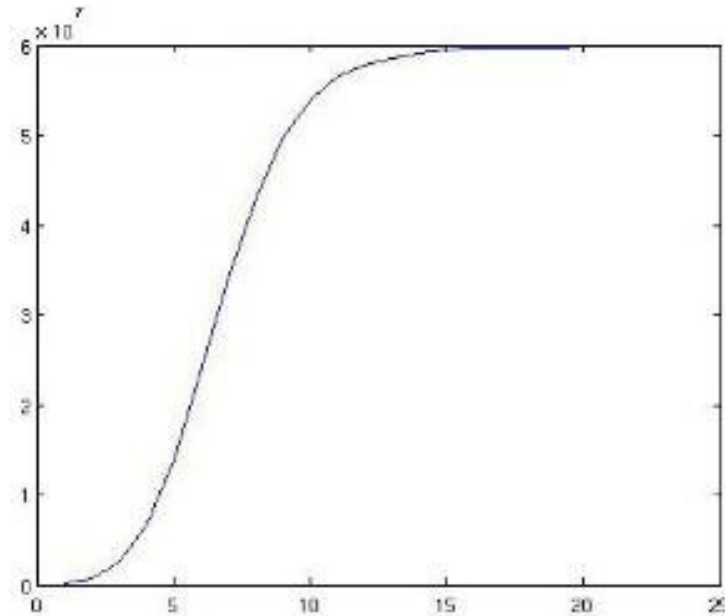
### 4 Results and Interpretation.

We present our results in the tables below:

**Table 1.** Results for indegree and outdegree exp. coefficients for the power law equation.

Sample	Number of Nodes	Indegree ex. Coeff. (k)			Outdegree ex. Coeff.(k)		
		I1	I2	I3	I1	I2	I3
Sample1	100K	2.17	2.16	2.16	2.23	2.21	2.18
Sample2	80K	2.09	2.06	2.04	2.11	2.07	2.07
Sample3	70K	2.11	2.08	2.06	2.15	2.13	2.12
Sample4	50K	2.06	2.06	2.05	2.12	2.11	2.11
Sample5	30K	2.13	2.13	2.10	2.13	2.12	2.12

Table 1 shows the coefficient of the power law for indegree and outdegree. I1, I2 and I3 represents the iteration level of SUBDUE. We see that the sample graphs and their subgraphs adhere to the power law, which is a property of Scale free Graph. And the value of the Degree coefficient is near about 2.1, conforming to the earlier calculations [2, 3].



**Fig. 1.** Graph diameter Vs Number of nodes

Figure 1 shows the graph for Diameter Vs Number of nodes in the graph. Diameter was calculated for the sample graphs and their respective subgraphs. We can see it follows nearly a logarithmic increment. For the Web graph of 100K nodes it was 18. The diameter saturates near the value 17. This again reaffirms the scale free nature of the web.

Table 2 gives the range of values of parameters for all the samples' three iterative subgraphs, formed after the compression in SUBDUE, gave the following values for:

**Table 2.** Graph Parameter values.

Parameter	Range
Clustering Coefficient	0.13-0.15
Betweenness Centralization	0.03-0.04
Closeness Centralization	0.36-0.41

The range of value of clustering index clearly indicates that the samples' subgraphs are much clustered and, so, exhibit small world network. The value range of Betweenness Centralization indicates that there are a few nodes in the sample graphs that lie in the path of most of the pairs and, hence have a significant influence over the communication of the other nodes. This indicates the presence of a core inside the subgraphs. The value range of closeness centralization indicates that there are a few nodes that are close to most of the other nodes in the subgraphs and, thus again, indicates towards the existence of a core in the sample graph.

Another interesting parameter that we evaluated was the compression quotient after every iteration level of SUBDUE.

$$Q = (V_f + E_f) / (V_i + E_i)$$

Where,  $V_f$  and  $E_f$  are number of Vertices and Edges in the compressed graph and  $V_i$  and  $E_i$  are the respective numbers in the original graphs.

**Table 3.** Results for compression Quotient

Sample	Number of nodes	Q1	Q2	Q3
Sample1	100K	0.931	0.847	0.863
Sample2	80K	0.891	0.871	0.875
Sample3	70K	0.953	0.941	0.949
Sample4	50K	0.924	0.911	0.918
Sample5	30K	0.956	0.892	0.922

We notice that the compression is more at the last level of SUBDUE iteration, which also supports the scale invariance in the sample webgraphs.

## 5 Conclusion.

We have carried out a pure graph based analysis of the web. And we have concluded from an entirely structural point of view that the Web is a fractal - It has cohesive sub-regions, at various scales, which exhibit the similar characteristics as the web for a lot of parameters. Each isomorphic subgraph nearly follows the classical Bow-Tie structure, with a robust core. This scalefree structural self similarity in the Web holds the key to building the theoretical models for understanding the evolution of the World Wide Web [2]. And further, this knowledge can be exploited while addressing the issues like security and routing measures for data streams, searching the internet and also e-marketing.

## References

1. R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, and E. Upfal, "Stochastic models for the web graph", In Proc. 41st FOCS, pages 57–65, 2000.
2. S. Dill, R. Kumar, K.S. Mccurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins, "Self-similarity in the web", ACM Trans. Inter. Tech., 2(3):205--223, 2002.
3. A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener, "Graph structure in the web", In Proc. 9th WWW, pages 309–320, 2000.
4. S. Chakarbarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan and S. Rajagopalan, "Automatic resource compilation by analyzing hyperlink structure and associated text", In Proceedings of the 7th WWW/Comput. Netw. 30, 1–7, 65–74, 1998.

5. S. Chakrabarti, B. Dom, D. Gibson, S. Ravi Kumar, P. Raghavan, S. Rajagopalan and A Tomkins, "Experiments in topic distillation", In SIGIRWorkshop on Hypertext Information Retrieval on the Web, 1998.
6. K.Bharat and M. Henzinger, "Improved algorithms for topic distillation in hyperlinked environments", In Proceedings of the 21st SIGIR, 104–111, 1998.
7. R.A. Botafogo and B. Shneiderman, "Identifying aggregates in hypertext structures", In Proceedings of the 3rd Hypertext Conference, 63–74, 1991.
8. J.Carriere, and R. Kazman, "WebQuery: Searching and visualizing the web through connectivity", In Proceedings of the 6th WWW 29, 8–13, 1257–1267, 1997.
9. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Trawling the web for cyber communities", In Proceedings of the 8th WWW/Comput. Netw. 31, 11-16, 1481–1493, 1999.
10. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Extracting large scale knowledge bases from the web", In Proceedings of the Conference on Very Large Data Bases, 639–650, 1999.
11. F. Harary, "Graph Theory", Addison Wesley, 1975.
12. L. Li, D. Alderson, R. Tanaka, J.C. Doyle, W. Willinger, "Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications", 2005.
13. M. Kuramochi, and G. Karypis, "Discovering Frequent Subgraphs", 2001
14. M. Kuramochi, and G. Karypis, "Discovering Frequent Geometric Subgraphs", 2002
15. X. Yan and J. Han, "gSpan: Graph-Based Substructure Pattern Mining", 2002.
16. M. Kuramochi, and G. Karypis, "GREW—A Scalable Frequent Subgraph Discovery Algorithm", 2003.
17. Cook Holder, et. al., "Subdue: Compression-based Frequent Pattern Discovery in Graph Data", Proceedings of the ACM KDD Workshop on Open-Source Data Mining, 2005.
18. P. Grunwald, "Tutorial Introduction to the Minimum Description Length Principle".
19. S. Suman and S. Aggarwal, "WebMine: A tool to uncover the web", 2005.