Computer Networks 33 (2000) 309–320

www.elsevier.com/locate/comnet

Graph structure in the Web

Andrei Broder

Keywords: Graph structure; Diameter; Web measurement

1. Introduction

Consider the directed graph whose nodes correspond to static pages on the Web, and whose arcs correspond to links between these pages. We study various properties of this graph including its diameter, degree distributions, connected components, and macroscopic structure. There are several reasons for developing an understanding of this graph.

(1) Designing crawl strategies on the Web [15].
(2) Understanding of the sociology of content creation on the Web.
(3) Analyzing the behavior of Web algorithms that make use of link information [9–11,20,26]. To take just one example, what can be said of the distribution and evolution of PageRank [9] values on graphs like the Web?

*

that a node has in-degree i is proportional to $1/i^x$, for some $x > 1$.

1.2.1. Zipf–Pareto–Yule and power laws

Distributions with an inverse polynomial tail have been observed in a number of contexts. The earliest observations are due to Pareto [27] in the context of economic models. Subsequently, these statistical behaviors have been observed in the context of literary vocabulary [32], sociological models [33], and even oligonucleotide sequences [24] among others.

ing several hundred million nodes, and a few billion arcs. We will refer to this graph as the *Web graph*, and our goal in this paper is to understand some of its properties. Before presenting our model for Web-like graphs, we begin with a brief primer on graph theory, and a discussion of graph models in general.

1.3. A brief primer on graphs and terminology

The reader familiar with basic notions from graph theory may skip this primer.

A *directed graph* consists of a set of *nodes*, denoted V and a set of *arcs*, denoted E. Each arc is an ordered pair of nodes $(u;v)$ representing a directed connection from u to v. The *out-degree* of a node u is the number of distinct arcs $.u;v_1/:::.u;v_k/$ (i.e., the number of links from u), and the *in-degree* is the number of distinct arcs $.v_1;u/:::.v_k;u/$ (i.e., the number of links to u). A path from node u to node v is a sequence of arcs $.u;u_1/;.u_1;u_2/;::::u_k;v/$. One can follow such a sequence of arcs to 'walk' through the graph from u to v. Note that a path from u to v does not imply a path from v to u. The *distance* from u to v is one more than the smallest k for which such a path exists. If no path exists, the distance from u to v is defined to be infinity. If $(u;v)$ is an arc, then the distance from u to v is 1.

Given a directed graph, a *strongly connected component* (strong component for brevity) of this graph is a set of nodes such that for any pair of nodes u and v in the set there is a path from u to v. In general, a directed graph may have one or many strong components. The strong components of a graph consist of disjoint sets of nodes. One focus of our studies will be in understanding the distribution of the sizes of strong components on the Web graph.

An *undirected graph* consists of a set of nodes and a set of *edges*, each of which is an unordered pair $\{u;v\}$ of nodes. In our context, we say there is an edge between u and v if there is a link between u and v, without regard to whether the link points from u to v or the other way around. The *degree* of a node u is the number of edges incident to u. A path is defined as for directed graphs, except that now the existence of a path from u to v implies a path from v to u. A *component* of an undirected graph is a set of nodes such that for any pair of nodes u and v ina 4 1 0 T [(a

2. Experiments and results

2.1. Infrastructure

All experiments were run using the Connectivity Server 2 (CS2) software built at Compaq Systems Research Center using data provided by AltaVista. CS2 provides fast access to linkage information on the Web. A build of CS2 takes a Web crawl as input and creates a representation of the entire Web graph induced by the pages in the crawl, in the form of a database that consists of all URLs that were crawled together with all in-links and out-links among those

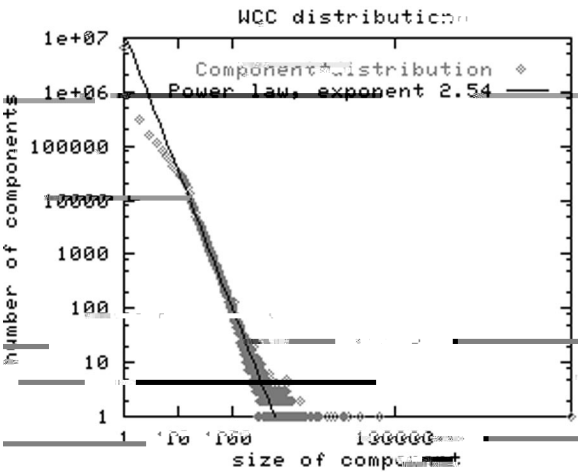


Fig. 5. Distribution of weakly connected components on the Web. The sizes of these components also follow a power law.

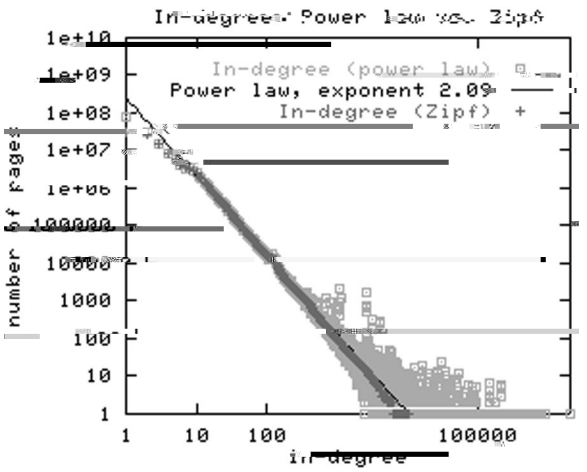
forward and backward directed links, the Web is a very well connected graph. Surprisingly, even the

this set has fewer than 90 nodes; in extreme cases it has a few hundred thousand), or it would 'explode' to cover about 100 million nodes (but never the entire 186 million). Further, for a fraction of the starting nodes, both the forward and the backward BFS runs would 'explode', each covering about 100 million nodes (though not the same 100 million in the two runs). As we show below, these are the starting points that lie in the SCC.

The cumulative distributions of the nodes covered in these BFS runs are summarized in Fig. 7. They reveal that the true structure of the Web graph must be somewhat subtler than a 'small world' phenomenon in which a browser can pass from any Web page to any other with a few clicks. We explicate this structure in Section 3.

2.2.5. Zipf distributions vs power law distributions

The *Zipf distribution* is an inverse polynomial function of *ranks* rather than magnitudes; for example, if only in-degrees 1, 4, and 5 occurred then a pow8(S)6(ec(e)-489(n)rec(e)-4(um]TJ TD o-8(t)-dde)-8(b0(b)1(a3956(v)13(e)-8(rse)-538(38(Sse14(w)129i)-1376ynomi)-5


In-degree: Power law vs. Zipf
In-degree (power law)
Power law, exponent 2.09
In-degree (Zipf)
number of pages
in-degree
1e+10
1e+09
1e+08
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
100000

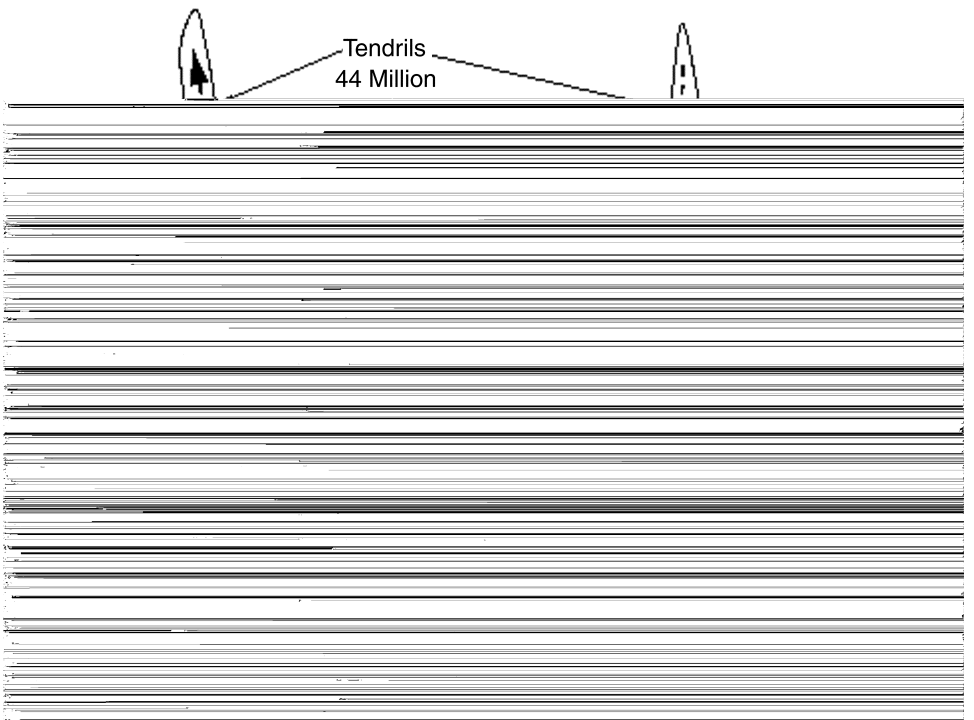


Fig. 9. Connectivity of the Web: one can pass from any node of IN through SCC to any node of OUT. Hanging off IN and OUT are TENDRILS containing nodes that are reachable from portions of IN, or that can reach portions of OUT, without passage through SCC. It is possible for a TENDRIL hanging off from IN to be hooked into a TENDRIL leading into OUT, forming a TUBE: i.e., a passage from a portion of IN to a portion of OUT without touching SCC.

regions have, if we explore in the direction ‘away’ from the center? The results are shown below in the row labeled ‘exploring outward – all nodes’.

Similarly, we know that if we explore in-links from a node in OUT, or out-links from a node in IN, we will encounter about 100 million other nodes in the BFS. Nonetheless, it is reasonable to ask: how many other nodes will we encounter? That is, starting from OUT (or IN), and following in-links

long path will be the same no matter which node of

(2) Mathematical models for evolving graphs, motivated in part by the structure of the Web; in addition, one may consider the applicability of such models to other large directed graphs such as the phone-call graph, purchase=transaction graphs, etc. [3].
(3) What notions of connectivity (besides weak and