Modeling the Webgraph Evolution with Graph Grammars

L. Ribeiro [a], L. Buriol [a], F.L. Dotti [b], C. Nunes [b], R. da Silva [a]

[a] *Instituto de Informática*
Universidade Federal do Rio Grande do Sul, Brazil
{*leila,buriol,rdasilva*}*@inf.ufrgs.br*

[b] *Faculdade de Informática*
Pontifícia Universidade Católica do Rio Grande do Sul, Brazil
{*fldotti,cnunes*}*@inf.pucrs.br*

The webgraph is the graph generated from the link structure of the web pages. In this graph, each node represents a web page and each edge is a hyperlink from one page to another. The webgraph has a particular structure, that is not similar to other known graphs, and it grows continuously. Nowadays it contains more then 25 billion nodes and about 400 billion edges. Even some properties of this graph have changed over time, there are some basic characteristics that remain the same. One of them is the power law distribution on the indegree of nodes. By this distribution, the probability that a node u be connected by k other nodes is about $k^{-\beta}$, that is, $Pr_u[IN(u) = k] \sim \frac{1}{k^\beta}$. The usual value of β in webgraphs is $\beta \approx 2.1$. Surprisingly, the power law distribution is found when analysing other properties of this graph. In 2000, Broder et al. [2] showed that the indegree and outdegree of the webgraph follows the power law distribution. In 2002, Pandurangan et al. [4] observed that the pagerank also follows the power law distribution. In 2004, Donato et al. [3] reported the power law on the distribution of the strongest connected component and their components. Based on some of these empirical observations in realworld webgraphs, models for generating synthetic webgraphs were proposed in the last few years. In all of them, the power law on the indegree distribution is the first property to be verified. A survey on models for generating webgraphs can be found in [1].

Graph grammars provide a formal way to generate graphs, based on the definition of the rules that govern the evolution of the graphs. In this paper, we investigate the suitability of graph grammars to generate and analyze the webgraph. The idea is to use properties that are observed in webgraphs and create rules that preserve these properties. That is, we create a grammar that

Preprint submitted to Elsevier Preprint

explains the evolution of the webgraph (concerning the given properties). As a first step, we will consider the power law distribution of the indegree of nodes, and show how to obtain a grammar that generates graphs preserving this property. We discuss the requirements on grammars and their semantics needed to faithfully model the evolution of the webgraph.

References

[1] Anthony Bonato. A survey on models of the webgraph. *Computer Networks*, pages 159–172, 2004.

[2] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, S. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33:309–320, June 2000.

[3] D. Donato, L. Laura, S. Leonardi, and S. Millozzi. Large scale properties of the webgraph. *European Physical Journal B*, 38:239–243, March 2004.

[4] G. Pandurangan, P. Raghavan, and E. Upfal. Using pagerank to characterize web structure. *Proc. of the 8th Annual International Conference on Combinatorics and Computing (COCOON)*, pages 330–339, 2002.