

# The WebGraph Framework II: Codes For The World-Wide Web

Paolo Boldi      Sebastiano Vigna

Dipartimento di Scienze dell'Informazione, Università degli Studi di Milano

A fundamental observation about compression of the web graph was made in the construction of the LINK database [3]: if we order URLs lexicographically, ordered successor lists tend to have small gaps, which can be coded using standard methods from full-text index construction.

To this purpose, we propose  $\zeta$  codes, a family of simple flat codes that generalise Elias'  $\gamma$  and are targeted at those gaps, which turn out to have a power-law distribution [2] with small exponent (around 1.2).

**Definition:** Given a fixed positive integer  $k$ , the *shrinking factor*, a positive integer  $x$  in the interval  $[2^{hk}, 2^{(h+1)k} - 1]$  is  $\zeta_k$ -coded by writing  $h + 1$  in unary (i.e., by writing  $h$  zeroes and a one), followed by a minimal binary coding of  $x - 2^{hk}$  in the interval  $[0, 2^{(h+1)k} - 2^{hk} - 1]$  (the minimal binary coding of  $x$  in the interval  $[0, z - 1]$  is the  $x$ -th binary word of length  $s - 1$  if  $x < 2^s - z$ , or the  $(x - z + 2^s)$ -th binary word of length  $s$  otherwise, where  $s = \lceil \log z \rceil$ ).

In the full version, we give the first thorough mathematical comparative analysis of several codes against power-law distributions: our analysis shows that  $\zeta$  codes should perform better (have smaller expected length) than both  $\delta$  and  $\gamma$  codes for values of the power-law exponent  $\alpha$  typically found in the gaps of web graphs, as shown in the table. The theoretical analysis has been confirmed by experimenting the codes against large datasets (about 1 Glink). The experiments were carried out using WebGraph [1], a framework that provides simple methods to manage very large graphs, specially tailored around web graphs. WebGraph contains a fully-documented implementation of various instantaneous codes, and of parameterisable compression algorithms that achieve the best compression rates known so far (about 2-3 bits per link for both the web graph and its transpose), exploiting both referentiation *à la* LINK and introducing a new technique called intervalisation; WebGraph also contains algorithms for accessing a compressed graph without actually decompressing it, using lazy techniques that delay the decompression until it is actually necessary.

$\alpha$	Best code
$< 1.06$	$\delta$
$[1.06, 1.08]$	$\zeta_6$
$[1.08, 1.11]$	$\zeta_5$
$[1.11, 1.16]$	$\zeta_4$
$[1.16, 1.27]$	$\zeta_3$
$[1.27, 1.57]$	$\zeta_2$
$[1.57, 2]$	$\gamma = \zeta_1$

- [1] Paolo Boldi and Sebastiano Vigna. The WebGraph framework I/II. Technical Reports 293-03/294-03, Università di Milano, Dipartimento di Scienze dell'Informazione, 2003. Available at <http://webgraph.dsi.unimi.it/>.
- [2] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web: experiments and models. In *Proc. of the 9th International World-Wide Web Conference*, 2000. Journal version in *Comput. Netw.* 33, 309.
- [3] K. Randall, R. Stata, R. Wickremesinghe, and J. Wiener. The LINK database: Fast access to graphs of the Web. Research Report 175, Compaq Systems Research Center, Palo Alto, CA, 2001.