2007
Volume 1, Number 2

ACM Transactions on

The Web

Association for Computing Machinery

Advancing Computing as a Science & Profession

ACM Transactions on the Web

Vol. 1 • No. 2 • 2007

Articles 7-10

ACM Transactions on
The Web

ACM
2 Penn Plaza, Suite 701
New York, NY 10121-0701
Tel.: (212) 869-7440
Fax: (212) 869-0481

Home Page: http://www.acm.org/tweb/

ACM Transactions on the Web (ISSN: 1559-1131) is published four times a year by the Association for Computing Machinery (ACM), 2 Penn Plaza, Suite 701, New York, NY 10121-0701. Periodicals class postage pending at New York, NY 10001, and at additional mailing offices. Printed in the U.S.A. POSTMASTER: Send address changes to *ACM Transactions on the Web*, ACM, 2 Penn Plaza, Suite 701, New York, NY 10121-0701.

For manuscript submissions, subscription, and change of address information, see inside backcover.

Association for
Computing Machinery

Advancing Computing as a Science & Profession

ACM Transactions on the Web

http://www.acm.org/tweb/

Guide to Manuscript Submission
Submission to the *ACM Transactions on the Web* is done electronically through http://acm.manuscriptcentral.com. Once you are at that site, you can create an account and password with which you can enter the ACM Manuscript Central manuscript review tracking system. From a drop-down list of journals, choose *Transactions on the Web* and proceed to the Author Center to submit your manuscript and your accompanying files.

You will be asked to create an abstract that will be used throughout the system as a synopsis of your paper. You will also be asked to classify your submission using the ACM Computing Classification System through a link provided at the Author Center. For completeness, please select at least one primary-level classification followed by two secondary-level classifications. To make the process easier, you may cut and paste from the list. Remember, you, the author, know best which area and sub-areas are covered by your paper; in addition to clarifying the area where your paper belongs, classification often helps in quickly identifying suitable reviewers for your paper. So it is important that you provide as thorough a classification of your paper as possible.

The ACM Production Department prefers that your manuscript be prepared in either LaTeX or Ms Word format. Style files for manuscript preparation can be obtained at the following location: http://www.acm.org/pubs/submissions/submission.htm. For editorial review, the manuscript should be submitted as a PDF or Postscript file. Accompanying material can be in any number of text or image formats, as well as software/documentation bundles in zip or tar-gzipped formats.

Questions regarding editorial review process should be directed to the Editor-in-Chief. Questions regarding the post-acceptance production process should be addressed to the Managing Editor, Roma Simon, at simon@hq.acm.org.

Subscription, Single Copy, and Membership Information.

Send orders to:

ACM
P.O. Box 12114
Church Street Station
New York, NY 10257

For information, contact:

Mail:	ACM Member Services Dept. 2 Penn Plaza, Suite 701 New York, NY 10121-0701
Phone:	+1-212-626-0500
Fax:	+1-212-944-1318
Email:	acmhelp@acm.org
Catalog:	http://www.acm.org/catalog

Subscription rates for *ACM Transactions on the Web* are $40 per year for ACM members, $35 for students, and $140 for nonmembers. Single copies are $18 each for ACM members and $40 for nonmembers. Your subscription expiration date is coded in four digits at the top of your mailing label; the first two digits show the year, the last two show the month of expiration.

Notice to Past Authors of ACM-Published Articles. ACM intends to create a complete electronic archive of all articles and/or other materials previously published by ACM. If you have written a work that was previously published by ACM in any journal or conference proceedings prior to 1978, or any SIG Newsletter at any time, and you do NOT want this work to appear in the ACM Digital Library, please inform permission@acm.org, stating the title of the work, the author(s), and where and when published.

Microfilm and Microfiche. Microfilm and microfiche editions are also available from University Microfilms International, 300 North Zeeb Road, Department PR, Ann Arbor, MI 48106.

About ACM. ACM, the Association for Computing Machinery, is an international scientific and educational organization dedicated to advancing the art, science, engineering, and application of information technology, serving both the professional and public interests by fostering the open interchange of information and by promoting the highest professional and ethical standards. In addition to *ACM Transactions on the Web*. ACM publishes numerous other refereed journals, magazines, newsletters, conference proceedings.

Visit ACM's Website: http://www.acm.org.

Change of Address Notification. To notify ACM of a change of address, use the addresses above or send an email to coa@acm.org.

Please allow 6–8 weeks for new membership or change of name and address to become effective. Send your old label with your new address notification. To avoid interruption of service, notify your local post office before change of residence. For a fee, the post office will forward 2nd- and 3rd-class periodicals.

Decoding the Structure of the WWW: A Comparative Analysis of Web Crawls

M. ÁNGELES SERRANO
Indiana University and Institute for Scientific Interchange, Turin, Italy
ANA MAGUITMAN
Universidad Nacional del Sur and CONICET
MARIÁN BOGUÑÁ
Universitat de Barcelona
and
SANTO FORTUNATO and ALESSANDRO VESPIGNANI
Indiana University and Institute for Scientific Interchange, Turin, Italy

The understanding of the immense and intricate topological structure of the World Wide Web (WWW) is a major scientific and technological challenge. This has been recently tackled by characterizing the properties of its representative graphs, in which vertices and directed edges are identified with Web pages and hyperlinks, respectively. Data gathered in large-scale crawls have been analyzed by several groups resulting in a general picture of the WWW that encompasses many of the complex properties typical of rapidly evolving networks. In this article, we report a detailed statistical analysis of the topological properties of four different WWW graphs obtained with different crawlers. We find that, despite the very large size of the samples, the statistical measures characterizing these graphs differ quantitatively, and in some cases qualitatively, depending on the domain analyzed and the crawl used for gathering the data. This spurs the issue of the presence of sampling biases and structural differences of Web crawls that might induce properties not representative of the actual global underlying graph. In short, the stability of the widely accepted statistical description of the Web is called into question. In order to provide a more accurate characterization of the Web graph, we study statistical measures beyond the degree distribution, such as degree-degree correlation functions or the statistics of reciprocal connections. The latter appears to enclose the relevant correlations of the WWW graph and carry most of the topological

This work was funded in part by the Spanish government's DEGS Grant No. FIS2004-05923-CO2-02 to M. B., by a Volkswagen Foundation grant to S. F., by NSF award 0513650 to A. C., and by the Indiana University School of Informatics.
Authors' addresses: M. A. Serrano (contact author), S. Fortunato, and A. Vespignani, School of Informatics, Indiana University, Bloomington, IN 47406; email: mariangeles.serrano@epfl.ch; fortunato@isi.it; alexv@indiana.edu; A. Maguitman, Universidad Nacional del Sur, 8000 Bahía Blanca, Argentina; email: agm@cs.uns.edu.ar; M. Boguñá, Departament de Física Fonamental, Universitat de Barcelona, 08028 Barcelona, Spain; email: marian.borguna@ub.edu.

information of the Web. The analysis of this quantity is also of major interest in relation to the navigability and searchability of the Web.

Categories and Subject Descriptors: H.4.m [**Information Systems Applications**]: Miscellaneous; G.3 [**Probability and Statistics**]

General Terms: Measurement

Additional Key Words and Phrases: Web graph structure, Web measurement, crawler biases, statistical analysis

ACM Reference Format:

Serrano, M. Á., Maguitman, A., Borguñá, M., Fortunato, S., and Vespignani, A. 2007. Decoding the structure of the WWW: A comparative analysis of Web crawls. ACM Trans. Web. 1, 2, Article 10 (August 2007), 25 pages. DOI = 10.1145/1255438.1255442 http://doi.acm.org/10.1145/1255438.1255442

1. INTRODUCTION

The World Wide Web (WWW) has grown at an unprecedented pace. While it is not possible to estimate its size because of pages with dynamical content, recent measures [Gulli and Signorini 2005] for the publicly indexable Web [Lawrence and Giles 1998; Lawrence and Giles 1999] point to the existence of many more than 10^{10} pages at the end of January 2005. Furthermore, the Web growth lacks any regulation and physical constraint (contrary to what happens with the physical Internet infrastructure [Pastor-Satorras and Vespignani 2004]), with new documents being added or becoming obsolete very quickly.

A fundamental step in decoding and understanding the WWW organization consists in the experimental studies of the WWW graph structure in which vertices and directed edges are identified with Web pages and hyperlinks, respectively. These studies are based on crawlers that explore the WWW connectivity by following the links on each discovered page, thus reconstructing the topology of the representative graph. Several studies based on those graphs have been performed in order to reveal the large-scale topological properties of the WWW. Distributions of in-degrees and out-degrees have been found to exhibit heavy-tails and the macroscopic architecture of connected components has made evident a rich structural organization, that is, the so-called bow-tie structure [Kumar et al. 1999; Barabási and Albert 1999; Barabási et al. 2000; Broder et al. 2000; Kumar et al. 2000; Adamic and Huberman 2001; Donato et al. 2004]. Reciprocal links and transitive relations regarding thematic communities [Eckmann and Moses 2002] have attracted attention as well, giving rise to a generally accepted picture of the topological structure of the WWW.

While the importance of these studies is indisputable, the dynamical nature of the Web and its huge size make very difficult the process of compressing, ranking, indexing, or mining the Web. Indeed, even the largest-scale Web crawlers cover only a small portion of the publicly available information and the obtained samples depend on the crawling policy employed [Cothey 2004]. In other words, it has been impossible so far to achieve any complete unbiased large-scale picture of the Web. On the other hand, the very large sizes of the gathered data sets have led to the general belief that the structural and statistical

properties observed in the WWW graphs were representative of the actual ones, thus leaving almost untouched the study of possible sampling biases [Henzinger et al. 2000; Bar-Yossef et al. 2000; Rusmevichientong et al. 2001; Cothey 2004]. In this respect, on the one hand it is crucial to get clear information about the exact policies and strategies followed by crawl engines, and, on the other hand, to explore to which extent the Web properties we observe are not biased by the specific characteristics of the crawls.

In this article, our primary objective is to stir discussion on the reliability of the widely accepted large-scale statistical properties of the Web and, at the same time, to provide new measures to discover whether or not inconsistencies are found when measuring the same properties across different crawls. To this end, we study four different data sets obtained in different years with different crawls and for different domains of the WWW. Web crawling policies are not taken into account so that we cannot evaluate the specific biases produced by crawlers. However, our results imply that we do not have yet an indisputable description of the large-scale structure of the Web and, as a consequence, the peculiarities of the artifacts that we use to explore it could be distorting its image. Our main contributions are as follows:

— We provide a careful comparative analysis of the structural and statistical large-scale topological properties of the different Web graphs, making evident qualitative and quantitative differences across different samples. We introduce higher-order statistical indicators characterizing single and two-vertex correlations in order to provide a full account of the connectivity pattern and structural ordering of the Web graph. We focus on correlations, which play a central discriminant role in model validation and can deeply affect the structure of the connected components of the graph. See Sections 4 and 5.
— We identify a crucial topological element, the reciprocal link, playing a key role in the organization of the WWW and accounting for most of the statistical correlations observed in Web graphs. Reciprocal links [Garlaschelli and Loffredo 2004], also referred to in the literature as *bidirectional links* [Boguñá and Serrano 2005] or colinks [Eckmann and Moses 2002], provide structural information that might be essential to assess how the underlying topology could affect the functionality [Boguñá and Serrano 2005] of the Web and/or processes running on it. Indeed, navigability and searchability are intimately related to the functionality of the WWW, and those properties strongly depend on the communication patterns among the constituent sites of the network. See Section 6.

2. RELATED WORK

The first empirical topological studies of the Web as a directed graph focused on the measure of the directed degree distributions $P(k_{in})$ and $P(k_{out})$, where the in/out-degree, k_{in} or k_{out} respectively, is defined as the number of incoming/outgoing links connecting a page to its neighbors. The work by Kumar et al. [1999] on a big crawl of about 40M nodes, and that by Barabási and Albert [1999] on a smaller set of over 0.3M nodes restricted to the domain of the University

of Notre Dame, suggested a scale-free nature for the WWW with power-law behaviors both for the in- and out-degree distributions.

Immediately after, a more complete investigation was published by Broder et al. [2000]. There, two sets from AltaVista crawls were analyzed, corresponding to different months in the same year 1999, May and October. The sets had over 200 million pages and 1.5 billion links. The authors reported detailed measurements on local and global properties of the Web graph which covered, for instance, the degree distributions, corroborating earlier observations, and also the presence and organization of connected components, unfolding the so-called bow-tie structure of the Web. One of the most intriguing conclusions there was that, from the analysis of those two sets, the observed structure of the Web was relatively insensitive to the particular large crawl used. In addition, the connectivity structure of the Web was resilient to the removal of a significant number of nodes.

Successively, further work [Donato et al. 2004] along the same lines has been performed over a large 2001 data set of 200M pages and about 1.4 billion edges made available by the WebBase project at Stanford (see next section for references and a project description). In this work, new measures were introduced along with the standard statistical observables, and the obtained results were compared with the ones presented in the work by Broder et al. [2000]. One of the reported differences is the deviation from the power-law behavior of the out-degree distribution.

On the other hand, the question whether subsets of the Web display the same characteristics as the Web at large has been discussed by a number of authors. Dill et al. [2001] found self-similarity within thematically unified subgraphs extracted from a single crawl of 60M pages gathered in October 2000. On the contrary, the different components of the bow-tie decomposition have been found to lack self-similarity in their inner structure when compared to the whole graph [Donato et al. 2005].

3. DATA SETS

We have analyzed and compared four sets of data corresponding to different years, from 2001 to 2004, and different domains, general and *.uk* and *.it* domains. The sets have been gathered within two different projects: the WebBase project and the WebGraph project, each using its own Web crawler, WebVac and UbiCrawler, respectively. *The WebBase Project* is a World Wide Web repository built as part of the Stanford Digital Libraries Project by the Stanford University InfoLab.[1] The Stanford WebBase project[2] [Hirai et al. 2000] is investigating various issues in crawling, storage, indexing, and querying of large collections of Web pages. The project aims to build the necessary infrastructure to facilitate the development and testing of new algorithms for clustering, searching, mining, and classification of Web content. The Stanford WebBase has been collected by the spider WebVac [Cho and Garcia-Molina 2000; Arasu et al. 2001] and makes available a Web repository with access to general crawls,

[1] `http://www-db.stanford.edu/.`

[2] `http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/.`

Table I. Number of Nodes and Edges of the Networks Considered, After Extracting Multiple Links and Self-Connections

Data set	WBGC01	WGUK02	WBGC03	WGIT04
# nodes	80,571,247	18,520,486	49,296,313	41,291,594
# links	752,527,660	292,243,663	1,185,396,953	1,135,718,909

such as the ones used in this research, or specific domain crawls restricted, for instance, to universities or institutions. *The WebGraph Project*[3] [Boldi and Vigna 2004] is being developed by the Laboratory for Web Algorithmics[4] (LAW) at the University of Milano and analyzes data obtained by its own crawler, UbiCrawler[5] [Boldi et al. 2004], designed to achieve high scalability and to be tolerant to failures.

The above projects provide several data sets publicly available to researchers. We analyze four samples ranging from 2001 to 2004. The WebBase general crawl of 2001 (WBGC01) and the WebBase general crawl of 2003 (WBGC03)[6] have been collected by the WebBase project in a general crawl using the WebVac spider. The remaining two sets collected by the UbiCrawler project, the WebGraph *.uk* domain of 2002 (WGUK02)[7] and WebGraph *.it* domain of 2004 (WGIT04),[8] are restricted to the domains *.uk* and *.it*, respectively. Note that the two domain crawls present an interesting difference. While pages in the *.uk* domain have higher probability to point to pages outside the domain, due to English being the official language in other influential countries, such as the U.S., and to the widespread use of English, the links in the Italian *.it* domain may be much more endogenous, which could potentially have a high effect on the Web description derived from the data.

We have cleaned the four sets by disregarding multiple links between the same pages and self-connections. In Table I we present a summary of the size in vertices and directed edges of the four sets analyzed in this article.

The following measures have been carried out using Matlab code.[9]

4. STRUCTURAL PROPERTIES

Data gathered in large scale crawls [Kumar et al. 1999; Barabási and Albert 1999; Barabási et al. 2000; Broder et al. 2000; Eckmann and Moses 2002; Donato et al. 2004] have uncovered the presence of a complex architecture underlying the structure of the Web graph. A widespread feature is the small-world property. Despite its huge size, the average number of URL links that must be followed to navigate from one document to the other, technically the average shortest path length, seems to be very small as compared to the value for a

[3]http://webgraph.dsi.unimi.it/.
[4]http://law.dsi.unimi.it/.
[5]http://ubi.iit.cnr.it/projects/ubicrawler/.
[6]ftp://db.stanford.edu/pub/webbase/.
[7]http://webdata.iit.cnr.it/unitedkingdom-2002/.
[8]http://webdata.iit.cnr.it/italy-2004/.
[9]Available upon request.

regular lattice of comparable size, and it seems to grow with the system size very slowly at a logarithmic pace [Albert et al. 1999; Broder et al. 2000]. Another important result is that the WWW exhibits a power-law relationship between the frequency of vertices and their degree, defined as the number of directed edges linking each vertex to its neighbors. This last feature is the signature of a very complex and heterogeneous topology with statistical fluctuations extending over many length scales [Albert et al. 1999; Barabási and Albert 1999; Kumar et al. 1999]. Finally, a fascinating macroscopic description of the Web has been provided by the study of the connected components, taking into account the directed nature of the Web graph [Broder et al. 2000]. In the following, we perform a careful comparative analysis of the four Web crawls described in the previous section. This will allow us to critically examine the stability of the various results as a function of the crawl and discuss which properties appear to be stable features of the global Web graph.

4.1 Sizes of Connected Components

The directed nature of the Web brings out a complex structure of connected components [Pastor-Satorras and Vespignani 2004; Dorogovtsev and Mendes 2003] that has been captured in the famous bow-tie architecture highlighted in the study presented in Broder et al. [2000]. If we disregard the directedness of links, the weakly connected component of the graph is made by all pages belonging to the giant component of the corresponding undirected graph. The undirected component becomes internally structured when the directed nature of the connections is considered. The most important of these new internal components is called the *strongly connected component* (SCC), which includes all pages mutually connected by a directed path. The other two relevant components are the *in-component* (IN) and the *out-component* (OUT). The first is formed by the vertices from which it is possible to reach the SCC by means of a directed path. The second refers to the set of vertices that can be reached from the SCC by means of a directed path. Finally, other secondary structures can also be present, such as tendrils, which contain pages that cannot reach the SCC and cannot be reached from it, or tubes which can directly connect the IN and OUT components without crossing the SCC. This complex composition is usually called the *bow-tie structure* because of the typical shape assumed by the figure sketching the relative size of each component (see Figure 1). It is clear that such a component structure is extremely relevant in the discussion of the functionalities of the Web. For instance, the relative sizes of the SCC and the IN and OUT components give us information about the probabilities of returning to an original page after exploration, or the size of the accessible Web once a starting page has been selected. The size of the SCC is of particular importance, since it constitutes the subset of reversible and complete access navigability. When one starts to surf the Web from the IN component, it is very likely that after a while one ends up in the SCC, and maybe eventually in the OUT component, but can never go back to the original point. Once in the OUT component, one can never go back to the other main components. But within the SCC, all nodes are reachable and can be revisited.

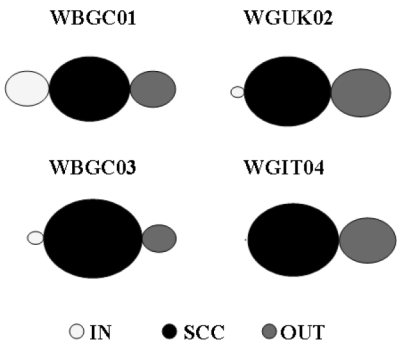

Fig. 1. Graphical representation of the sizes of the global components reported in Table II. The area of each component is proportional to its actual size in number of nodes, so that the relative sizes of the components in the figure account for the actual relative sizes of the Web graphs.

Table II. Sizes of the SCC, IN, and OUT Components and Their Union MAIN, Which Does Not Contain Either Tendrils or Tubes, So That It Differs from the Weakly Connected Component (Values are shown as a percentage of the total number of nodes and as a percentage of the total number of links. Notice that for nodes MAIN=SCC+IN+OUT, but for links it does not hold since only internal connections are considered.)

Nodes	WBGC01	WGUK02	WBGC03	WGIT04
IN	17.24%	1.69%	2.28%	0.03%
SCC	56.46%	65.28%	85.87%	72.30%
OUT	17.94%	31.88%	11.26%	27.64%
MAIN	91.64%	98.85%	99.41%	99.98%
Links	WBGC01	WGUK02	WBGC03	WGIT04
IN	9.78%	1.30%	0.01%	0.01%
SCC	65.79%	77.87%	88.68%	81.57%
OUT	11.52%	14.11%	0.30%	12.39%
MAIN	95.36%	99.33%	99.90%	99.99%

We summarize the values for the sizes of the components of the four data sets in Table II. The figures for the domain crawls are in agreement to those reported in Donato et al. [2005], where the same *.uk* and *.it* sets were also examined. The analysis of the four data sets considered in the present study shows a noticeable variability of the basic component structure of the resulting graph. In particular, the IN component is the most unstable feature that ranges from accounting for about 20% of the total structure (WBGC01) to the case in which it is practically absent (WGIT04). This variability could be likely ascribed to the different crawling strategies and the fact that each of those may use different starting points. Moreover, crawlers perform a directed exploration in the sense that they follow outgoing hyperlinks to reach pointed pages, but cannot navigate backwards using incoming hyperlinks. This implies that the exploration of the IN component could be strongly biased by the initial conditions used to start

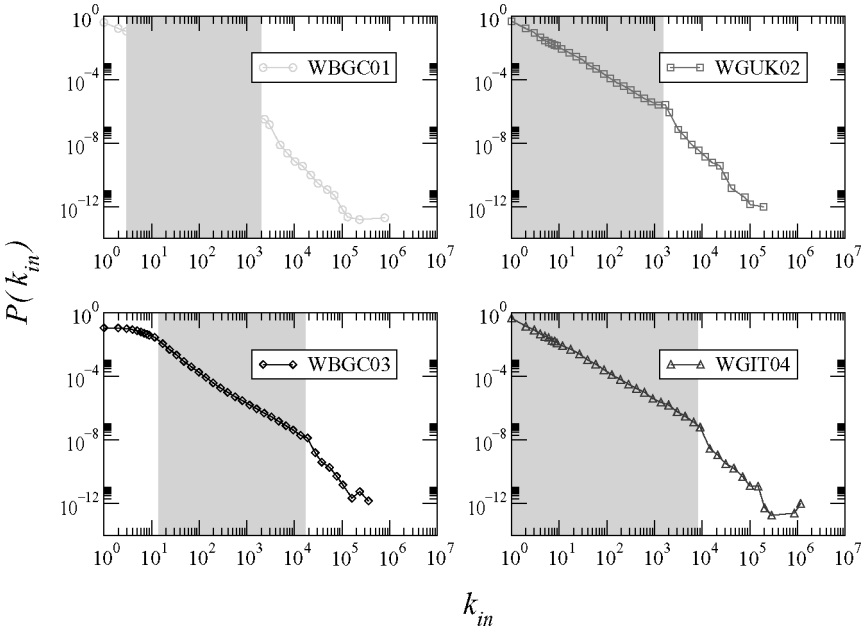

Fig. 2. Distributions of incoming links. In the shadowed regions all the functions decay as a power-law with exponents given in Table III.

the crawl. Variations are however not limited to the IN component. Also the relative sizes of the SCC and the OUT component vary from sample to sample, even by a factor close to three in the case of the OUT component. Finally, notice that the sizes of the IN and OUT components of the WBGC01 set are quite symmetric, as was also found in Broder et al. [2000], where the values reported for the sizes of the IN, SCC, and OUT of components of the AltaVista crawl were 21.3%, 27.7%, and 21.2%, respectively. In summary, it is evident from this analysis that the structure of Web graphs is strongly dependent on the data set considered.

4.2 Degree Distributions

A major interesting feature found in Web graphs is the presence of a highly heterogeneous topology, with degree distributions characterized by wide variability and heavy tails [Albert et al. 1999; Barabási and Albert 1999; Kumar et al. 1999]. The degree distribution $P(k)$ for undirected networks is defined as the probability that a node is connected to k other nodes. For directed networks, this function splits in two separate functions, the in-degree distribution $P(k_{in})$ and the out-degree distribution $P(k_{out})$, which are measured separately as the probabilities of having k_{in} incoming links and k_{out} outgoing links, respectively. In Figures 2 and 3 we report the behavior of the in-degree and out-degree distributions. These distributions, as for most real world networks, are found to be very different from the degree distribution of a random graph or an ordered lattice. They are both skewed and spanning several orders of magnitude in degree

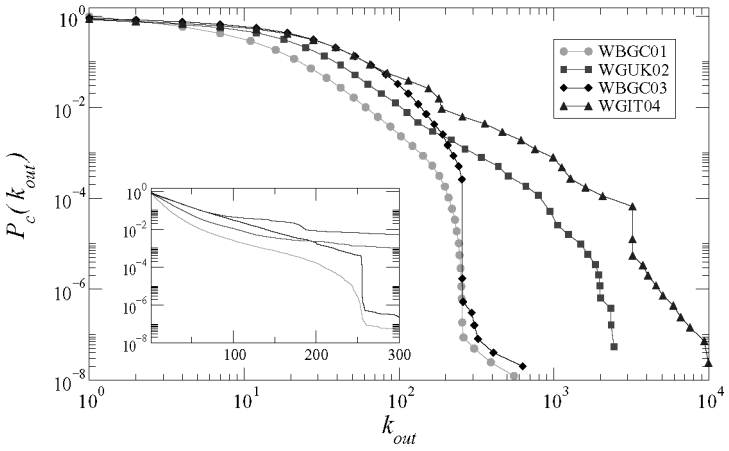

Fig. 3. Distributions of outgoing links. For visualization purposes, we use cumulative distributions defined as $P_c(k_{out}) = \sum_{k'_{out} \geq k_{out}} P(k'_{out})$. The inset shows the same curves in a linear-log scale.

values. The in-degree distribution exhibits a heavy-tailed form approximated by a power-law behavior $P(k_{in}) \sim k_{in}^{-\gamma_{in}}$, generally spanning over three to four orders of magnitude. In Figure 2, we show the region considered in the evaluation of the exponent obtained by a maximum likelihood algorithm for discrete distributions. The in-degree distributions also exhibit a noisy tail that cannot be well fitted with a specific analytic form. Yet it strengthens the evidence for the heavy-tailed character of $P(k_{in})$.

A different situation is faced in the case of the out-degree distribution $P(k_{out})$. In this case, a clear exponential cutoff is observed and the range of degree values is two to four orders of magnitude smaller than what is found for the in-degree distribution. The origin of the cutoff can be explained by the different nature of the in-degree and out-degree evolution. The in-degree of a vertex is the sum of all the hyperlinks incoming from all the Web pages in the WWW. In principle, thus, there is no limit to the number of incoming hyperlinks, that is determined only by the popularity of the Web page itself. On the contrary, the out-degree is determined by the number of hyperlinks present in the page, which are controlled by Web administrators. For evident reasons (clarity, handling, data storage), it is very unlikely to find an excessively large number of hyperlinks in a given page. This represents a sort of finite capacity [Mossa et al. 2002] for the formation of outgoing hyperlinks that might naturally lead to a finite cutoff in the out-degree distribution.

The heavy-tailed behavior of the in-degree distribution implies that there is a statistically significant probability that a vertex has a very large number of connections compared to the average degree $\langle k_{in} \rangle$. In addition, the extremely large value of $\langle k_{in}^2 \rangle$, and therefore of the variance $\sigma^2 = \langle k_{in}^2 \rangle - \langle k_{in} \rangle^2$, is signaling the extreme heterogeneity of the connectivity pattern, since it implies that statistical fluctuations are virtually unbounded, and tells us that the average degree is not the typical degree value in the system, that is, we have scale-free distributions. The heavy-tailed nature of the degree distribution

Table III. Main Statistical Properties of the Analyzed Sets: Average Degree $\langle k \rangle$, Maximum Degree k_{max}, Standard Deviation σ, Heterogeneity Parameter κ, and Maximum Likelihood Estimate of the Exponent of the Power-law In-Degree Distribution γ_{in} (Precision Error $\pm$ 0.1) (All values are provided for in- and out-degrees and for the four data sets. The symbol ∞ for γ_{out} means that the out-degree distributions decay faster than a power-law.)

Data set	WBGC01	WGUK02	WBGC03	WGIT04
$\langle k_{in} \rangle$	9.3	15.8	24.1	27.5
k_{in}^{max}	788632	194942	378875	1326744
σ_{in}	200.2	143.3	421.6	881.4
κ_{in}	4298.6	1317.5	7414.9	28269.9
γ_{in}	1.9	1.7	2.2	1.6
	WBGC01	WGUK02	WBGC03	WGIT04
$\langle k_{out} \rangle$	9.3	15.8	24.1	27.5
k_{out}^{max}	552	2449	629	9964
σ_{out}	13.1	27.4	29.5	67.1
κ_{out}	27.7	63.4	60.3	191.0
γ_{out}	∞	∞	∞	∞

has also important consequences in the dynamics of processes taking place on top of these networks. Indeed, recent studies about network resilience in front of removal of vertices [Cohen et al. 2000] and spreading phenomena [Pastor-Satorras and Vespignani 2001] have shown that the relevant parameter for these phenomena is the ratio between the first two moments of the degree distribution $\kappa = \langle k^2 \rangle / \langle k \rangle$. If $\kappa \gg 1$, the network manifests some properties that are not observed for networks with exponentially decaying degree distributions. In the case of directed networks, this heterogeneity parameter has to be defined separately for in- and out-degrees as $\kappa_{in} = \langle k_{in}^2 \rangle / \langle k_{in} \rangle$ and $\kappa_{out} = \langle k_{out}^2 \rangle / \langle k_{out} \rangle$,[10] since it could happen that the network is heterogeneous with respect to one of the degrees but not to the other.[11] In Table III, we provide these values for the empirical graphs along with a summary of the numerical properties of the probability distributions analyzed so far. The heavy-tailed behavior is especially evident when comparing the heterogeneity parameters κ and their wide range variations. A marked difference is observed for the out-degree distributions where the variance and heterogeneity parameters are indicating a limited variability of the function $P(k_{out})$. From the exponents reported for the in-degree distribution, it clearly results that the fittings to a power-law form can yield different results, depending on the data set analyzed. These variations could signal a slightly different structure of the Web graph depending on the domain crawled or the eventual presence of statistical biases due to the

[10]Notice that for any directed graph $\langle k_{in} \rangle = \langle k_{out} \rangle$.

[11]In addition, a third parameter can be defined which accounts for the effect of the crossed one point correlations $\kappa_{in,out} = \langle k_{in} k_{out} \rangle / \langle k_{in} \rangle$.

crawling strategy. It is interesting to notice that a similar variability is encountered in studies of the power-law behavior of Web samples restricted to specific thematic groups [Pennock et al. 2002]. Another oddity that has to be signaled is the fact that the general crawls WBGC01 and WBGC03 exhibit a much smaller cutoff of the out-degree distribution than observed in the two domain crawls. This is somehow counterintuitive given the larger sizes of the general crawls. A possible explanation is that long pages surpassing a certain length limit may have been truncated. This might hint at the presence of a bias in the way hyperlinks are explored by different crawlers.

5. DEGREE CORRELATIONS

As an initial discriminant of structural ordering, the attention has been focused on the networks' degree distribution. This function is, however, only one of the many statistics characterizing the structural and hierarchical ordering of a network. A full account of the connectivity pattern calls for the detailed study of degree correlations, which has not been done so far. Beyond their theoretical relevance, correlations affect, in some cases in a dramatic way, percolation processes and, in particular, the structure of the connected components [Boguñá and Serrano 2005]. More specifically, the presence or absence of local correlations between in- and out-degree can determine whether the network will or will not have a bow-tie structure despite the degree distribution remaining unchanged. In general, correlations are critical in the diffusion, navigation and spreading processes on complex networks. Their structure and robustness as well are heavily affected, and community identification algorithms or hierarchical ordering are also altered or determined by degree correlation properties. For instance, in the context of the physical Internet recent works have proposed that topology metrics are generally interdependent and that the structure of the correlation functions defines the metrics that underly the topology of the network [Mahadevan et al. 2006]. Correlations are also important as a decisive discriminant in model validation. Along these lines, a quantitative study of the mixing properties of networks through local correlation measures and opportune projection of the degree-degree joint probability distribution is fundamental.

5.1 Single Vertex Degree Correlations

First, we examine local one-point degree correlations for individual nodes, in order to understand if there is a relation between the number of incoming and outgoing links in single pages. Since most of the analyzed degree distributions are heavy-tailed, fluctuations are extremely large so that the linear correlation coefficient is not well defined for those cases. Instead, we provide the crossed one-point correlations, $\langle k_{in} k_{out} \rangle$, normalized by the corresponding uncorrelated value, $\langle k_{in} \rangle \langle k_{out} \rangle$. We also report the function

$$\langle k_{out}(k_{in}) \rangle = \frac{1}{N_{k_{in}}} \sum_{i \in \Upsilon(k_{in})} k_{out,i}, \tag{1}$$

Table IV. Crossed In-Degree Out-Degree Correlations for Individual Nodes, Normalized by the Uncorrelated Values

Data set	WBGC01	WGUK02	WBGC03	WGIT04
$\frac{\langle k_{in}k_{out}\rangle}{\langle k_{in}\rangle\langle k_{out}\rangle}$	2.8	3.1	1.6	5.6

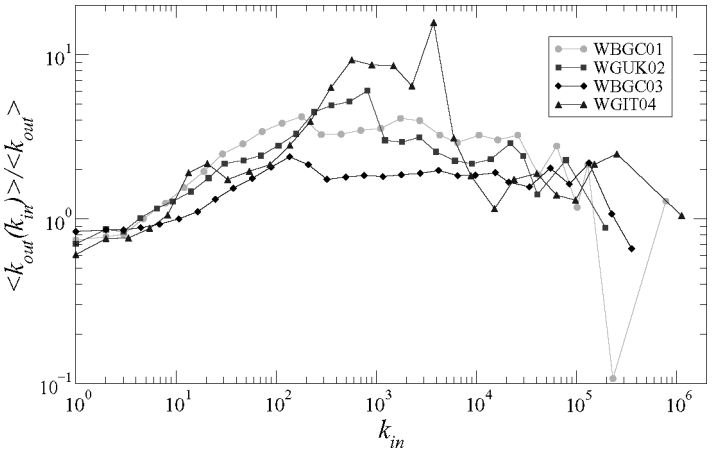

Fig. 4. Normalized average out-degree as a function of the in-degree for the four different data sets.

which measures the average out-degree of nodes as a function of their in-degree. $N_{k_{in}}$ stands for the number of nodes with in-degree k_{in} and $k_{out,i}$ is the out-degree of node i. The notation $i \in \Upsilon(k_{in})$ indicates that the summation has to be performed over the set of nodes of degree k_{in}, denoted by $\Upsilon(k_{in})$. The results can be found in Table IV and in Figure 4.

A significant positive correlation between the in-degrees and the out-degrees of single nodes is found for all the sets. That means that more popular pages tend to point to a higher number of other pages. This positive correlation is found to be true for a range of in-degrees that spans from $k_{in} = 1$ to $k_{in} = 10^2 \sim 10^3$, depending on the specific set. Beyond this point no noticeable correlation is observed; see Figure 4. The set for the Italian domain is more noisy, but this pattern appears to be independent of the crawl used to gather the data and, thus, it seems to be a stable feature of the Web.

5.2 Two-Vertex Degree Correlations

Another important source of information about the network structural organization lies in the correlations of the degrees of neighboring vertices. These correlations can be probed in undirected networks by inspecting the average degree of nearest neighbors of a vertex i, where *nearest neighbors* refers to the set of vertices at a hop distance equal to 1,

$$\overline{k_{nn,i}} = \frac{1}{k_i} \sum_{j \epsilon \nu(i)} k_j . \tag{2}$$

The sum runs over the nearest-neighbor vertices of each vertex i, gathered in the set $\nu(i)$. From this quantity, a convenient measure is obtained by averaging over degree classes to obtain the average degree of the nearest neighbors for vertices of degree k, defined as [Pastor-Satorras et al. 2001]

$$\overline{k_{nn}}(k) = \frac{1}{N_k} \sum_{i \in \Upsilon(k)} \overline{k_{nn,i}} = \sum_{k'} k' P(k'|k), \tag{3}$$

where N_k is the number of nodes with degree k, the notation $i \in \Upsilon(k)$ indicates that the summation has to be performed over the set of nodes of degree k, denoted by $\Upsilon(k)$, and $P(k'|k)$ quantifies the conditional probability that a vertex with degree k is connected to a vertex with degree k'. This measure provides a sharp proof of the presence or absence of correlations. In the case of uncorrelated networks, the degrees of connected vertices are independent random quantities, so that $P(k'|k)$ is only a function of k'. In this case, $\overline{k_{nn}}(k)$ does not depend on k and equals $\kappa = \langle k^2 \rangle / \langle k \rangle$. Therefore, a function $\overline{k_{nn}}(k)$ showing any explicit dependence on k signals the presence of degree correlations in the system. Real networks usually tend to display one of two different patterns [Newman 2002]. Assortative networks exhibit $\overline{k_{nn}}(k)$ functions increasing with k, which denotes that vertices are preferentially connected to other vertices with similar degree. Examples of assortative behavior are typically found in many social structures. On the other hand, disassortative networks exhibit $\overline{k_{nn}}(k)$ functions decreasing with k, which denotes that vertices are preferentially connected to other vertices with very different degree. Examples of disassortative behavior are typically found in several technological networks, as well as in communication and biological networks.

In the case of the WWW, the study of the degree-degree correlation functions is naturally affected by the directed nature of the graph. In Barrat et al. [2004], a set of directed degree-degree correlation functions was defined considering that, in this case, the neighbors can be restricted to those connected by a certain type of directed link, either incoming or outgoing. For the WWW, we study the most significant distributions, taking into account that we can partition the neighborhood of each single node i into neighboring nodes connected to it by incoming links and neighboring nodes connected to it by outgoing links. A first correlation indicator, $\overline{k_{in,nn}}(k_{in})$, is defined as the normalized average in-degree of the neighbors of nodes of in-degree k_{in}, when those neighboring nodes are found following incoming links of the original node (see Figure 5(a)). If we measure the popularity of Web pages in terms of the number of pages pointing to them, this function quantifies the average popularity of pages pointing to pages with a certain popularity. The exact definition is given in the Appendix along with the expression for the normalization factor. The rest of the correlation functions, $\overline{k_{out,nn}}(k_{in})$, $\overline{k_{out,nn}}(k_{out})$, $\overline{k_{in,nn}}(k_{out})$, can be defined in an analogous manner. Each plot in Figure 6 shows these correlation functions for the four data sets analyzed in this article. Remarkably, only one of the functions shows an increasing pattern denoting the presence of assortative correlations for the four data sets. The average out-degree of neighbors of nodes of high out-degree is also high, so that the average number of references is high in pages pointed by pages

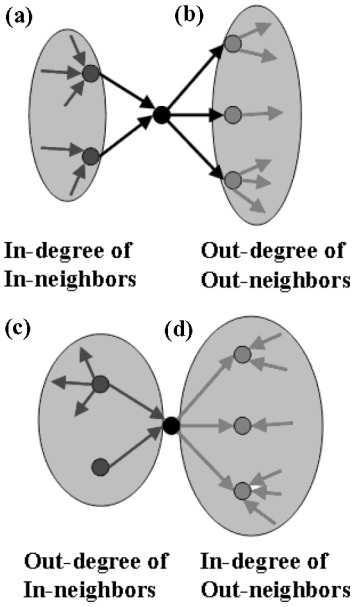

Fig. 5. Graphical sketch illustrating the degree-degree correlation functions defined in Section 5.2. We focus on a single node—the central node in the figures—with in-degree $k_{in} = 2$ and out-degree $k_{out} = 3$. In (a) the average in-degree of its in-neighbors is computed taking into account the incoming arrows inside the grey area. The function $\overline{k_{in,nn}}(k_{in})$ is then the average of this quantity over all nodes with the same in-degree. The rest of the functions are defined in a similar way, as highlighted in (b), (c), and (d).

with a high number of references. In all other cases, very mild or a complete lack of correlation is observed. For instance, this has important implications in the study of PageRank, which has been shown to be linearly correlated with the in-degree if degree-degree correlations are absent [Fortunato et al. 2006].

6. THE ROLE OF RECIPROCAL LINKS

While a directed network, the Web has many pages pointing to each other. A couple of pages pointing to each other corresponds to the presence of a reciprocal link that can be considered as undirected. These reciprocal connections play an important role as percolation catalysts, favoring the appearance of the fine structure (IN, OUT, and SC components) in the giant connected component [Boguñá and Serrano 2005] and potentially affect the navigability of the Web. Furthermore, they most probably denote relations at a peer level in contraposition to hierarchy, so that they could be central elements in community detection, for instance. In this section we introduce and investigate reciprocal links as crucial elements in the understanding of the WWW. To this end, we will differentiate into incoming, outgoing, and reciprocal links, where incoming and outgoing links do not include the ones taking part in reciprocal connections and are referred to as nonreciprocal. This allows us to consider reciprocal and nonreciprocal connections as separate and well-defined independent entities

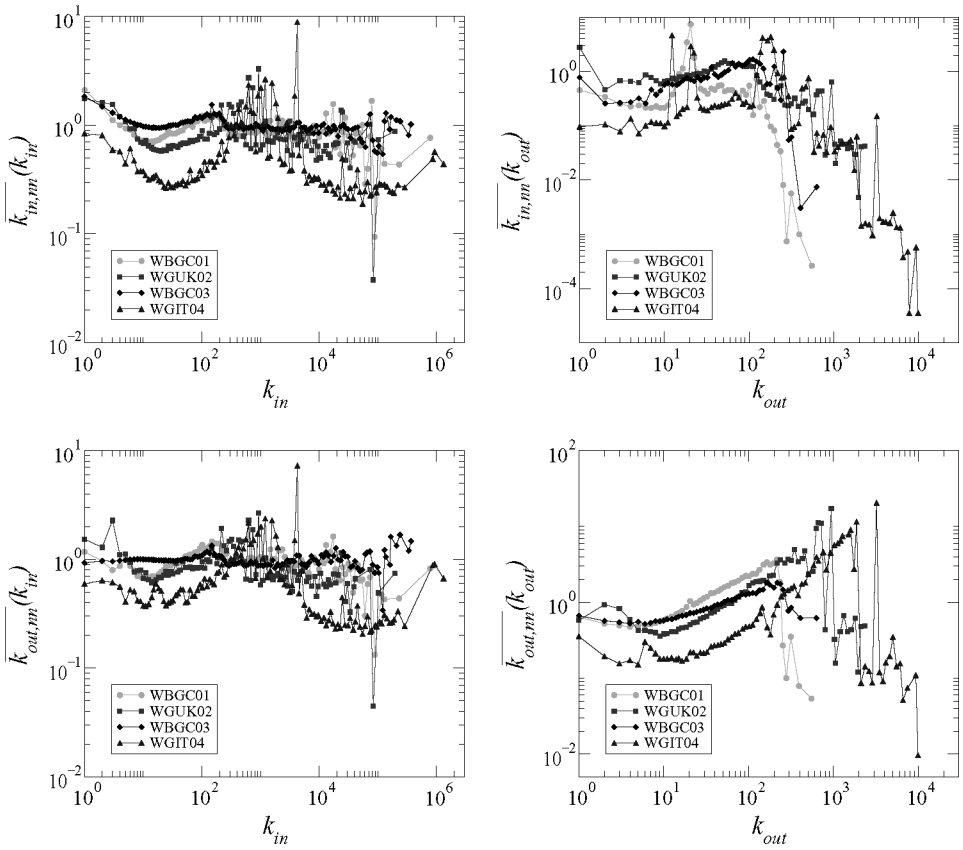

Fig. 6. Degree-degree correlations for the four different data sets. Explicit expressions for the quantitative definition of these functions can be found in the Appendix.

and to provide a statistical analysis able to capture additional information on the Web structure.

6.1 Degree Distributions

For the sake of notation, in the following we will identify the nonreciprocal in-degree and out-degree of a given vertex i with $q_{in,i}$ and $q_{out,i}$, respectively. Analogously, the reciprocal degree (r-degree) $q_{r,i}$ indicates the number of reciprocal connections to neighboring vertices. The degree distributions of nonreciprocal links are extremely similar to those obtained for the global in and out-degree. On the other hand, the reciprocal degree distribution appears to exhibit a strikingly different behavior depending on the crawl examined. In particular, general crawls show a distribution $P(q_r)$ with an exponentially fast decaying behavior, while the domain crawls have a heavy-tailed distribution varying over three orders of magnitude (see Figure 7). In Table V, we summarize the main properties of $P(q_r)$ for the various data sets. Also, from the values shown there one can easily see the mild fluctuations and heterogeneity expressed by the general crawl data sets. The evident differences in the

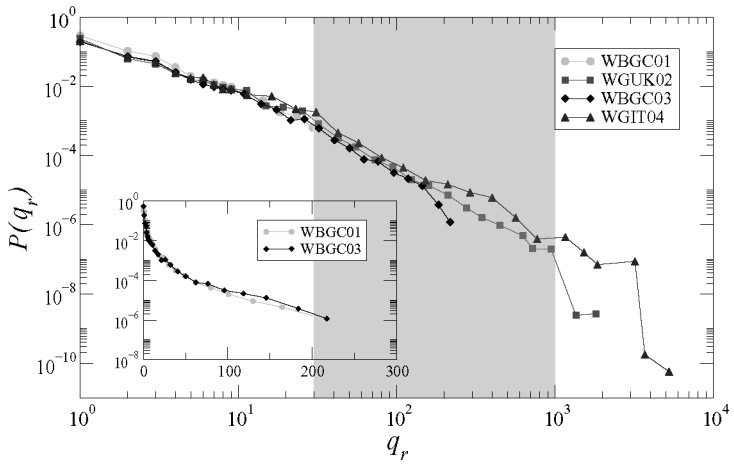

Fig. 7. Probability distributions of reciprocal links. The inset shows the distributions for the two general crawls in a linear-log scale.

Table V. Main Statistical Properties of the Reciprocal Subgraphs: Average Degree $\langle q_r \rangle$, Maximum Degree q_r^{max}, Standard Deviation σ_r, Heterogeneity Parameter κ_r, and Maximum Likelihood Estimate of the Exponent of the Power-Law in-Degree Distribution γ_r (Precision Error ± 0.1) (The symbol ∞ means that the distribution decays faster than a power-law.)

Data set	WBGC01	WGUK02	WBGC03	WGIT04
$\langle q_r \rangle$	2.7	3.3	2.4	5.2
q_r^{max}	391	1997	253	6164
σ_r	7.2	16.2	8.1	42.7
κ_r	21.9	82.7	30.0	352.6
γ_r	∞	2.6	∞	2.6

reciprocal degree distributions match the dissimilar component structure observed in general and domain crawls. On the other hand, the origin of the two different statistical behaviors does not find a clear explanation and deserves further investigation. Finally, once again we have to emphasize the odd finding of general crawls showing reciprocal degree distribution cutoffs much smaller than those observed for domain crawls, possibly due to a smaller threshold value for the length of the explored pages.

6.2 One-Point Degree Correlations

The distinction between reciprocal and nonreciprocal links induces a higher complexity even at the most local level. In this case, each node is characterized by three different quantities. Consequently, we need to introduce three correlation measures, that is, the average nonreciprocal out-degree as a function of the nonreciprocal in-degree, $\langle q_{out}(q_{in}) \rangle$, and the average r-degree as a

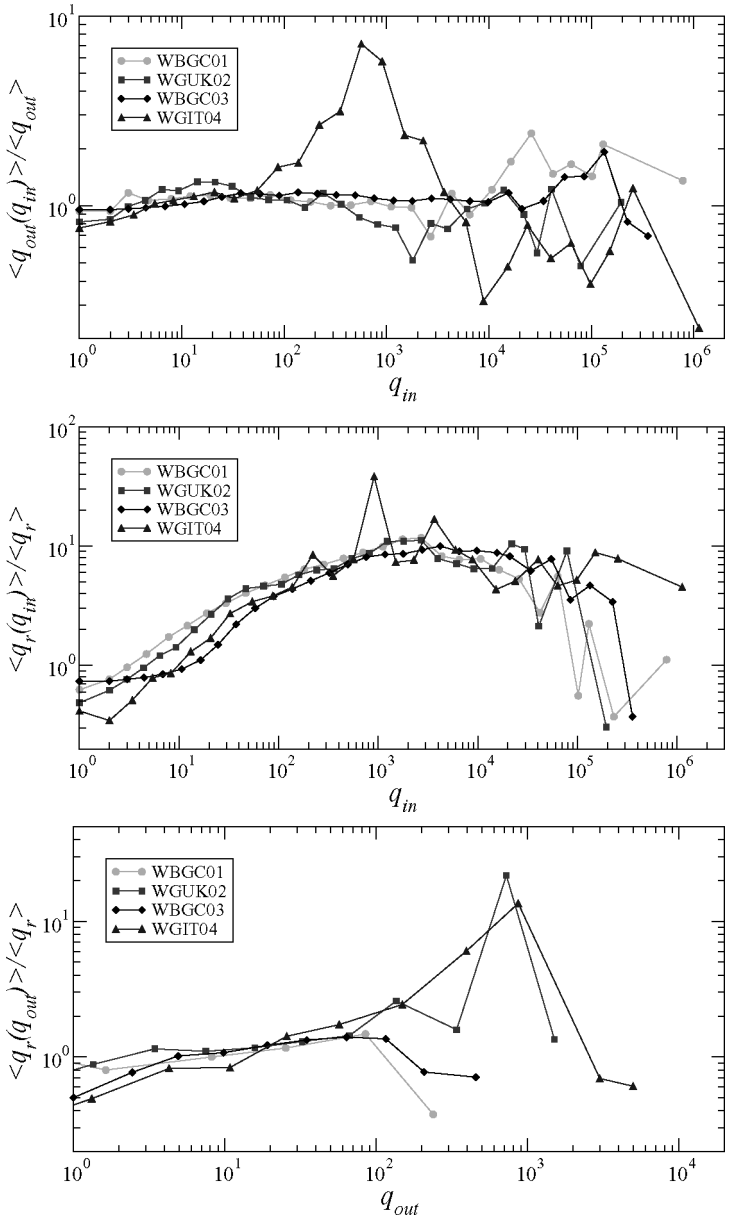

Fig. 8. One node correlations for the four different data sets. The functions shown are the normalized average nonreciprocal out-degree as a function of the nonreciprocal in-degree, and the normalized average r-degree as a function of the nonreciprocal in- and out-degrees.

function of the number of nonreciprocal incoming and outgoing links, $\langle q_r(q_{in})\rangle$ and $\langle q_r(q_{out})\rangle$, respectively (see Figure 8). A surprising result is that, in this case, there is no clear correlation between nonreciprocal in- and out- degrees but there is a positive correlation between reciprocal and nonreciprocal in-degrees;

Table VI. Crossed Nonreciprocal In-Degree, Out-Degree, and r-Degree Correlations for Individual Nodes

Data set	WBGC01	WGUK02	WBGC03	WGIT04
$\frac{\langle q_{in}q_{out}\rangle}{\langle q_{in}\rangle\langle q_{out}\rangle}$	1.0	0.9	1.1	2.0
$\frac{\langle q_{in}q_{r}\rangle}{\langle q_{in}\rangle\langle q_{r}\rangle}$	6.7	7.4	6.0	9.9
$\frac{\langle q_{out}q_{r}\rangle}{\langle q_{out}\rangle\langle q_{r}\rangle}$	1.1	1.4	1.3	2.4

see Table VI. So the positive correlation previously observed in Section 5.1 between in- and out-degrees is just a consequence of this new correlation between reciprocal and nonreciprocal in-degrees.

6.3 Degree-Degree Correlations

The two vertices correlation analysis presented in Section 5.2 can be repeated for the nonreciprocal and reciprocal decomposition of the network. Now, we have to differentiate reciprocal links and segregate the neighborhood of each single node i into neighboring nodes connected to it by nonreciprocal incoming links, neighboring nodes connected to it by nonreciprocal outgoing links, and neighboring nodes connected to it by reciprocal links. The degree-degree correlation functions corresponding to the first two cases give similar results to the ones presented in the previous section and do not signal the presence of any relevant correlation pattern (not plotted).

A very different picture is obtained when we measure correlations following reciprocal connections. A strong positive correlation is observed between the in-degrees of nodes connected by reciprocal links. This is clearly visible in the upper left plot of Figure 9, which shows the normalized average nonreciprocal in-degree of the neighbors of nodes of nonreciprocal in-degree q_{in}, when the neighbors are found following reciprocal links, $\overline{q_{in,nn}}(q_{in}|r)$. This function shows a clear increase of two orders of magnitude as a function of q_{in}, indicating an assortative correlation. The same behavior is found between nonreciprocal out-degrees (lower right plot of Figure 9). Concerning the crossed correlations, we observe again a positive correlation between the neighboring nonreciprocal in-degree and the nonreciprocal out-degree but no noticeable correlation in the opposite one, that is, the average nonreciprocal out-degree of the reciprocal neighbors of a node is independent of the nonreciprocal in-degree of that node (see lower left plot in Figure 9). In summary, the analysis of the two-vertex degree correlation behavior indicates that most of the structural correlations of Web graphs are found in vertices connected by reciprocal links. This type of links therefore represents an element of particular interest in that they express the ordering principles (beyond simple randomness) at the basis of the Web structure.

6.4 The Reciprocal Subgraph

Very interesting information is provided by the study of how reciprocal links are structurally organized among them. If we look at the subgraph formed by

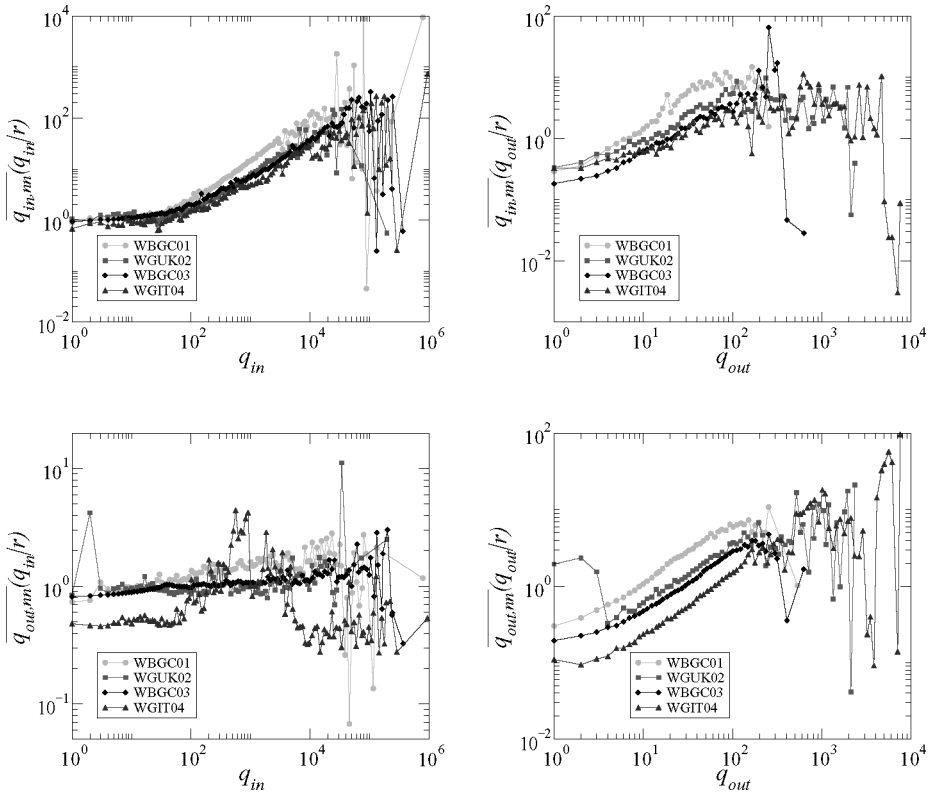

Fig. 9. Nonreciprocal degree-degree correlations for the four different data sets.

the vertices and the reciprocal links we can use the tools adopted for undirected graphs. A measure of the two vertices correlation function is therefore expressed by $\overline{q_{r,nn}}(q_r)$ (see Section 5.2), that is, the standard measure of an undirected network if we identify reciprocal links as undirected. As shown in Figure 10, this function shows a first decrease, for $q_r < 10$, followed by a linear increase up to a critical value depending on the crawler. At high reciprocal degrees, a cloud of points is populating the low r-degree region of the average nearest-neighbor reciprocal degree. This defines a bimodal pattern which indicates two different behaviors. The low values cloud can be interpreted as a collection of starlike structures, with central hubs connected to low degree nodes. The decrease for low degrees of the function $\overline{q_{r,nn}}(q_r)$ suggests that this behavior could not be produced by statistical fluctuations. High-degree nodes in the cloud are connected to low-degree nodes and, therefore, low-degree nodes should be connected to higher-degree nodes, as we indeed observe in the deviation from the linear behavior in the low-degree range. This effect is probably due to the "home" button in many Web pages that belong to a bigger site. The linear behavior may have two different interpretations. The first one is that the network is a tree in which high-degree nodes are connected to other high-degree nodes. The second one is that the network forms cliquelike structures, that is, groups of

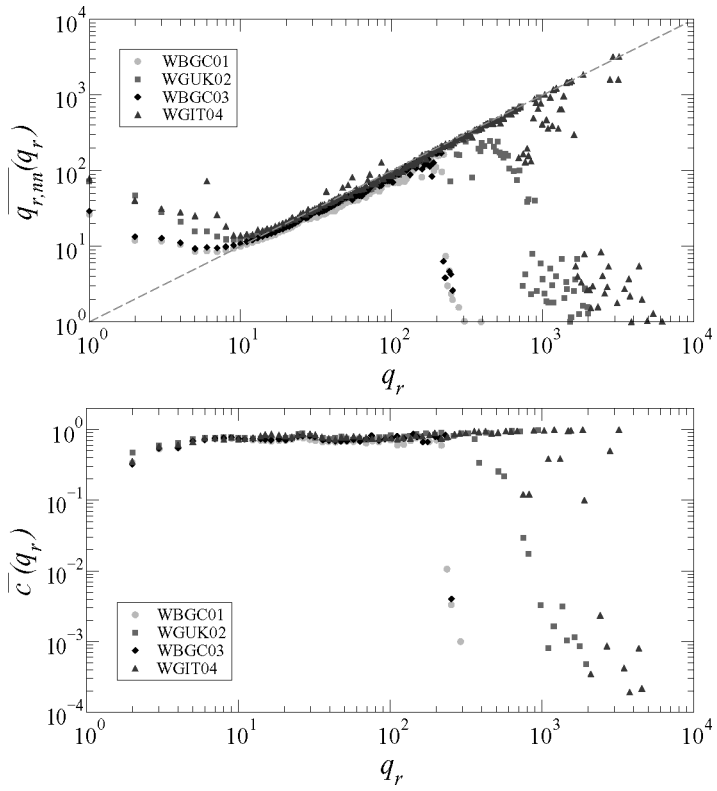

Fig. 10. Average nearest-neighbors degree (top) and degree-dependent clustering coefficient (bottom) for the reciprocal links and for all the samples.

pages pointing simultaneously to each other. To discern which scenario is more appropriate, we inspect the local connectivity properties of reciprocally linked vertices. Since we can treat the reciprocal subgraph as an undirected one, we can probe the local interconnectedness by analyzing the clustering coefficient defined as the fraction of interconnected neighbors of j: $c_j = 2 \cdot n_{\texttt{link}}/(q_{r,j}(q_{r,j}-1))$, where $n_{\texttt{link}}$ is the number of reciprocal links between the $q_{r,j}$ reciprocal neighbors of j. This quantity measures the density of interconnected vertex triplets and it is therefore close to 1 in the case of a fully interconnected neighborhood and zero in the case of a tree structure. Global statistical information can be gathered by inspecting the average clustering coefficient $\overline{c}(q_r)$ restricted to classes of vertices with reciprocal degree q_r. In the first scenario, $\overline{c}(q_r)$ should be very small and decreasing with the degree because of the treelike structure. In the second one, $\overline{c}(q_r)$ should be significant and independent of the degree. In Figure 10, we show the function $\overline{c}(q_r)$ which exhibits a high and constant value followed by a cloud of points with very low clustering coefficient at the same point where the function $\overline{q}_{r,nn}(q_r)$ also splits. This indicates that the organization of the reciprocal subgraph is a set of starlike structures combined with cliques, or communities, of highly interconnected pages. Very interestingly, this

Table VII. Level of Agreement That the Four Different Data Sets Show on the Properties Analyzed in This Article (GC stands for general crawl, DC for domain crawl, and NR for nonreciprocal)

Property	Agreement	Comments	Section
Sizes of the connected components	No	Strong differences	4.1
In-degree distribution	Yes	Heavy-tailed, but strong variation in exponents	4.2
Out-degree distribution	No	Exponential for GC, heavy-tailed for DC	4.2
Correlation between in-degree and out-degree	Yes	Significant	5.1
Degree-degree correlations	Yes	Not significant, except for $\overline{k_{out,nn}}(k_{out})$	5.2
Reciprocal degree distribution	No	Exponential for GC, heavy tailed for DC	6.1
Correlations between NR out-degree and NR in-degree or reciprocal degree	Yes	Not significant	6.2
Correlations between NR in-degree and reciprocal degree	Yes	Significant	6.2
Degree-degree correlations through reciprocal links	Yes	Significant, except for $\overline{q_{out,nn}}(q_{in}\|r)$	6.3
Degree-degree correlations reciprocal subgraph	Yes	Significant	6.4
Clustering reciprocal subgraph	Yes	Significant	6.4

pictorial characterization appears to be the same in all Web graphs considered, pointing to a stable feature of the Web graph. The present analysis identifies in the reciprocal subgraph an important element that might help in decoding the structure of the WWW. Finally, we have to stress that the reciprocal component is surely extremely important for the analysis and understanding of navigation patterns and the network resilience to link removal.

7. CONCLUSIONS

Contrary to what happened with the scrutiny of Internet maps, the issue of sampling biases in the structure of the WWW has been left almost untouched. The large size of the data sets has led to the belief that the global properties were well defined in view of the abundant statistics available. Noticeably, from

the present analysis, it appears that the resulting picture of the WWW structure and its statistical characterization could be considerably affected by the design of the tools we use to observe it. While some of the basic properties are qualitatively preserved across different data sets and stay as good candidates for genuine properties, other features and quantities are highly variable. In Table VII, we present a summary of the level of agreement that the different data sets show on the properties analyzed in this article. This results in a fuzzy picture of the WWW structure, where sampling biases may still play a major role. Our main conclusion is then that so far, and despite an approximate view of the Web from data provided by Web crawlers, we still lack an exact and definitive description of its large-scale properties and architecture, which could be affecting how effectively we can navigate, search, index, or mine it.

The present work thus highlights the need for a theoretical framework able to approach a detailed analysis and understanding of the sampling biases implicit in the most widely used crawling strategies. In this sense, numerical studies of simulated exploration of directed network models could be a starting point to approach this problem and to have a preliminary assessment of the intrinsic biases induced by the crawling process. Moreover, differences among crawls should be further investigated in relation to the crawling policies adopted in designing the engines. To this end, it is essential that a full and detailed report of the particular Web-crawling policies and strategies that have been pursued in obtaining the data is systematically provided along with the data sets.

Finally, the results presented in this article are potentially helpful for improving the design of future crawlers, not only regarding latent biases. These applications are improved to a great extent when they take advantage of the special hyperlink structure among Web documents. In this respect, correlations and reciprocal links could play a key role which has to be explored in more detail.

APPENDIX. DEGREE-DEGREE CORRELATIONS: QUANTITATIVE DEFINITIONS

We study the most significant two-point correlation functions, taking into account that we can segregate the neighborhood of each single node i into neighboring nodes connected to it by incoming links, the set $\nu_{in}(i)$, and neighboring nodes connected to it by outgoing links, the set $\nu_{out}(i)$. Following Equation (3), we can write

$$\begin{aligned}
\overline{k_{in,nn}}(k_{in}) &= \frac{1}{\kappa_{in,out}} \frac{1}{N_{k_{in}}} \sum_{i \in \Upsilon(k_{in})} \frac{\sum_{j \epsilon \nu_{in}(i)} k_{in,j}}{k_{in,i}}, \\
\overline{k_{out,nn}}(k_{in}) &= \frac{1}{\kappa_{out}} \frac{1}{N_{k_{in}}} \sum_{i \in \Upsilon(k_{in})} \frac{\sum_{j \epsilon \nu_{in}(i)} k_{out,j}}{k_{in,i}}, \\
\overline{k_{in,nn}}(k_{out}) &= \frac{1}{\kappa_{in}} \frac{1}{N_{k_{out}}} \sum_{i \in \Upsilon(k_{out})} \frac{\sum_{j \epsilon \nu_{out}(i)} k_{in,j}}{k_{out,i}}, \\
\overline{k_{out,nn}}(k_{out}) &= \frac{1}{\kappa_{in,out}} \frac{1}{N_{k_{out}}} \sum_{i \in \Upsilon(k_{out})} \frac{\sum_{j \epsilon \nu_{out}(i)} k_{out,j}}{k_{out,i}}.
\end{aligned} \tag{4}$$

These measures are normalized by the corresponding uncorrelated values defined in Section 4.2 as the heterogeneous parameters $\kappa_{in,out}$, κ_{in}, and κ_{out}, in order to make them independent of the system size and so comparable across samples.

The same quantities can be calculated when nonreciprocal and reciprocal links are differentiated. Now the neighborhood of each single node i is segregated into neighbors connected to it by nonreciprocal incoming links, the set $\nu_{in}^{nr}(i)$, neighbors connected to it by nonreciprocal outgoing links, the set $\nu_{out}^{nr}(i)$, and neighbors connected to it by reciprocal links, the set $\nu_r(i)$. The functions given in Equation (4) are valid whenever the in and out subscripts are restricted to nonreciprocal links. When following only reciprocal links, one can redefine them in a similar way:

$$\begin{aligned}
\overline{q_{in,nn}}(q_{in}|r) &= \frac{1}{\kappa_{r,in}} \frac{1}{N_{q_{in}}} \sum_{i \in \Upsilon(q_{in})} \frac{\sum_{j \epsilon \nu_r(i)} q_{in,j}}{q_{r,i}}, \\
\overline{q_{out,nn}}(q_{in}|r) &= \frac{1}{\kappa_{r,out}} \frac{1}{N_{q_{in}}} \sum_{i \in \Upsilon(q_{in})} \frac{\sum_{j \epsilon \nu_r(i)} q_{out,j}}{q_{r,i}}, \\
\overline{q_{in,nn}}(q_{out}|r) &= \frac{1}{\kappa_{r,in}} \frac{1}{N_{q_{out}}} \sum_{i \in \Upsilon(q_{out})} \frac{\sum_{j \epsilon \nu_r(i)} q_{in,j}}{q_{r,i}}, \\
\overline{q_{out,nn}}(q_{out}|r) &= \frac{1}{\kappa_{r,out}} \frac{1}{N_{q_{out}}} \sum_{i \in \Upsilon(q_{out})} \frac{\sum_{j \epsilon \nu_r(i)} q_{out,j}}{q_{r,i}};
\end{aligned} \tag{5}$$

the normalization terms in this case are

$$\begin{aligned}
\kappa_{r,in} &= \tfrac{\langle q_r q_{in} \rangle}{\langle q_r \rangle}, \\
\kappa_{r,out} &= \tfrac{\langle q_r q_{out} \rangle}{\langle q_r \rangle}.
\end{aligned} \tag{6}$$

ACKNOWLEDGMENTS

We acknowledge the Stanford WebBase project and the LAW WebGraph project for providing publicly available data. We would also like to thank Filippo Menczer for helpful discussions and valuable comments.

REFERENCES

Adamic, L. A. and Huberman, B. A. 2001. The Web's hidden order. *Commun. ACM 44*, 9, 55–60.

Albert, R., Jeong, H., and Barabási, A.-L. 1999. Diameter of the World-Wide Web. *Nature 401*, 6749, 130–131.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., and Raghavan, S. 2001. Searching the Web. *ACM Trans. Internet Tech. 1*, 1, 2–43.

Bar-Yossef, Z., Berg, A., Chien, S., Fakcharoenphol, J., and Weitz, D. 2000. Approximating aggregate queries about web pages via random walks. In *Proceedings of the 26th International Conference on Very Large Data Bases* (VLDB). 535–544.

Barabási, A.-L. and Albert, R. 1999. Emergence of scaling in random networks. *Science 286*, 5439, 509–512.

Barabási, A.-L., Albert, R., and Jeong, H. 2000. Scale-free characteristics of random networks: The topology of the World-Wide Web. *Physica A 281*, 1-4, 69–77.

Barrat, A., Barthélemy, M., and Vespignani, A. 2004. Traffic-driven model of the World Wide Web graph, Stephano Leonardi, Ed. *Algorithms and Models for the Web-Graph*. Lecture Notes in Computer Science, vol. 3243. Springer, Berlin, Heidelburg, Germany, 56–67.

Boguñá, M. and Serrano, M. A. 2005. Generalized percolation in random directed networks. *Phys. Rev. E 72*, 1, 016106.

Boldi, P., Codenotti, B., Santini, M., and Vigna, S. 2004. Ubicrawler: A scalable fully distributed Web crawler. *Softw. Pract. Exper. 34*, 8, 711–726.

Boldi, P. and Vigna, S. 2004. The Webgraph framework i: Compression techniques. In *WWW 2004 Conference Proceedings*. ACM, New York, NY, 595–601.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, S., Tomkins, A., and Wiener, J. 2000. Graph structure in the Web. *Comput. Netw. 33*, 1-6, 309–320.

Cho, J. and Garcia-Molina, H. 2000. The evolution of the Web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Databases* (Cairo, Egypt). 200–209.

Cohen, R., Erez, K., ben Avraham, D., and Havlin, S. 2000. Resilience of the Internet to random breakdown. *Phys. Rev. Lett. 85*, 21, 4626.

Cothey, V. 2004. Web-crawling reliability. *J. Amer. Soc. Inform. Sci. Techn. 55*, 14, 1228–1238.

Dill, S., Kumar, R., McCurley, K., Rajagopalan, S., Sivakumar, D., and Tomkins, A. 2001. Self-similarity in the Web. In *Proceedings of the 27th International Conference on Very Large Data Bases* (VLDB). 69–78.

Donato, D., Laura, L., Leonardi, S., and Millozzi, S. 2004. Large scale properties of the Webgraph. *Eur. Phys. J. B 38,* 2, 239–243.

Donato, D., Leonardi, S., Millozzi, S., and Tsaparas, P. 2005. Mining the inner structure of the Web graph. In *Proceedings of the Eighth International Workshop on the Web and Databases* (WebDB). 145–150.

Dorogovtsev, S. N. and Mendes, J. F. F. 2003. *Evolution of Networks: From Biological Nets to the Internet and WWW*. Oxford University Press, Oxford, U. K.

Eckmann, J. P. and Moses, E. 2002. Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Procc. Natl. Acad. Sci. 99*, 9, 5825–5829.

Fortunato, S., Boguñá, M., Flammini, A., and Menczer, F. 2006. Approximating pagerank from in-degree. In *cs.IR/0511016, presented at the Fourth Workshop on Algorithms and Models for the Web-Graph*, Nov. 30 – Dec. 1, Banff, Alta., (Canada).

Garlaschelli, D. and Loffredo, M. I. 2004. Patterns of link reciprocity in directed networks. *Phys. Rev. Lett. 93*, 26, 268701.

Gulli, A. and Signorini, A. 2005. The indexable Web is more than 11.5 billion pages. In *WWW 2005 Conference Proceedings* (Chiba, Japan). ACM Press, New York, NY, 902–903.

Henzinger, M. R., Heydon, A., Mitzenmacher, M., and Najork, M. 2000. On near-uniform URL sampling. In *WWW 2000 Conference Proceedings* (Amsterdam, The Netherlands). ACM Press, New York, NY, 295–308.

Hirai, J., Raghavan, S., Paepcke, A., and Garcia-Molina, H. 2000. Webbase: A repository of Web pages. In *WWW 2000 Conference Proceedings* (Amsterdam, The Netherlands). ACM Press, New York, NY, 277–293.

Kumar, R., Raghavan, P., Rajagopalan, S., Sivakumar, D., Tomkins, A., and Upfal, E. 2000. Stochastic models for the Web graph. In *Proceedings of the* 41*th IEEE Symposium on Foundations of Computer Science* (FOCS). 57–65.

Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. 1999. Trawling emerging cyber-communities automatically. In *WWW 1999 Conference Proceedings* (Toronto, Ont., Canada). ACM Press, New York, NY, 3–4.

Lawrence, S. and Giles, C. L. 1998. Searching the world wide web. *Science 280*, 5360, 98–100.

Lawrence, S. and Giles, C. L. 1999. Accessibility of information on the Web. *Nature 400*, 6740, 107–109.

Mahadevan, P., Krioukov, D., Fall, K., and Vahdat, A. 2006. Systematic topology analysis and generation using degree correlations. In *Proceedings of SIGCOMM06* (Pisa, Italy). ACM Press, New York, NY.

Mossa, S., Barthélemy, M., Stanley, H. E., and Amaral, L. A. N. 2002. Truncation of power law behavior in scale-free network models due to information filtering. *Phys. Rev. Lett. 88*, 13, 138701.

Newman, M. E. J. 2002. Assortative mixing in networks. *Phys. Rev. Lett. 89*, 20, 208701.

PASTOR-SATORRAS, R., VÁZQUEZ, A., AND VESPIGNANI, A. 2001. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett. 87*, 25, 258701.

PASTOR-SATORRAS, R. AND VESPIGNANI, A. 2001. Epidemic spreading in scale-free networks. *Phys. Rev. Lett. 86*, 14, 3200–3203.

PASTOR-SATORRAS, R. AND VESPIGNANI, A. 2004. *Evolution and Structure of the Internet. A Statistical Physics Approach*. Cambridge University Press, Cambridge, U. K.

PENNOCK, D. M., FLAKE, G. W., LAWRENCE, S., GLOVER, E. J., AND GILES, C. L. 2002. Winners don't take all: Characterizing the competition for links on the web. *Proc. Natl. Acad. Sci. 99*, 8, 5207–5211.

RUSMEVICHIENTONG, P., PENNOCK, D. M., LAWRENCE, S., AND GILES, C. L. 2001. Methods for sampling pages uniformly from the World Wide Web. In *Proceedings of the AAAI Fall Symposium on Using Uncertainty Within Computation*. 121–128.

Received March 2006; revised December 2006; accepted February 2007