Webmetrics

Mario A. Maggioni
Università Cattolica di Milano, Italy

Teodora Erika Uberti
Università Cattolica di Milano, Italy

THE INTERNET: A COMPLEX NETWORK

The Internet is perhaps one of the newest and most powerful media that enables the transmission of digital information and communication across the world, even if there still exist important divides (digital divide) between and within countries in the endowment, access and use of this technology.

To a certain extent, the level and rate of the Internet diffusion reflects its nature, being a complex structure subject to positive network externalities (which are at the basis of the so-called "Metcalfe's law," which states that the value of a network increases with the number of nodes that belong to it: the larger the number of nodes joining a network, the more valuable the network).

In addition, the Internet is a network that evolves dynamically over time; hence, it is important to define its nature, main characteristics and potentialities.

THE INTERNET AND THE WWW

To investigate the nature of the Internet, it is essential to distinguish between the physical infrastructure (which we will call "Internet") and its virtual, graphical and multimedia interface, the World Wide Web (WWW), a service platform which, on January 2004, was made of 4,606,743 pages (Zakon, 2004) and in 2003, its surface was estimated to be equal to 167 terabytes (Lyman & Varian, 2003).

The Internet is a series of connected networks; each of them is composed by a set of Internet hosts and computers connected via traditional or optical cables; while the WWW is constituted by Web pages and Web sites connected by Internet hyperlinks, enabling information and communication to flow from one computer to another. Therefore, the Internet is the physical infrastructure reflecting the technical capability of a given geographical area (i.e., a country, region or city) to enable effective and efficient exchanges of digital information; while the WWW is a virtual space reflecting the ability to create and exchange digital information and contents. Of course, the latter would not exist without the former (Abbate, 1999; Barners- Lee & Fischetti, 1999).

Both the Internet and the WWW are networks, but while the first has a relatively stable infrastructure (because investments to implement and maintain it are rather large and costly, and the subjects involved are a limited number: mostly corporate, governmental or non-governmental organizations), the second changes very rapidly over time (because is cheap and easy to create and maintain a Web site, and the number of people agents involved is huge); therefore, it is very difficult to give a precise and updated description of it.

The most common indicator of the Internet diffusion is the number of Internet hosts, that should reveal the ability of a given geographical area to create digital contents and support the exchange flow of information. Unfortunately, this definition is ambiguous, and its measurement does not entirely capture the actual diffusion of this medium. First, generic Top Level Domains (gTLDs; which account for almost 67% of the total hosts in January 2004) do not reflect any specific geographical location. Second, some country code top-level domains (ccTLDs), even if from a formal perspective, are unambiguously geo-located, display a mismatch between the location of the TLD and the actual source of digital information. For example, .tv domain (acronym for Tuvalu Island) is very diffused among televisions' corporate because of its abbreviation, and the digital contents are not related to this country. Similarly, .nu

(acronym for Niue Island) is quite common because of phonetic reasons, but not because Niue inhabitants frequently use the Internet. Third, even if considered jointly with other technological (i.e., number of computers, telephone lines) and economic indicators (GDP per capita), the number of Internet hosts may capture a large share of the Internet infrastructure, but misses mapping the flows of digital information.

Hence, it is crucial to use suitable indicators to map the infrastructure of flows of digital information and contents across the WWW. The number of Web pages and sites reflects the amount of information available on the WWW, but misses the description of the structure of digital information flows, the ability to create digital contents and to attract e-attention (not to mention the crucial issue of the quality of information).

WEBOMETRICS PROCEDURES USING INTERNET HYPERLINKS

A relevant problem in the analysis of the WWW concerns measurement. Almind and Ingwersen (1997), referring to the organizational nature of this service platform – a network of dynamically linked pages – adopted quantitative techniques, derived from bibliometric and infometric procedures, to analyze the structure and the use of information resources available on the WWW. Hence, they introduced the term Webometrics: the bibliometric study of Web pages.

The intuition of these authors was to adapt citations' analysis and quantitative analysis (i.e., impact factors) to the Web space to enable the investigation of Web pages' contents and to rank Web sites according to their use or "value" (calculated through hyperlinks acting as papers' citations); to allow the evaluation of WWW organizational structure; to study net surfers' Web usage and behavior; and finally, to check Web technologies (i.e., retrieval algorithms adopted by different search engines).

The starting point of Webometrics is taking into account the structure of the WWW: a network of Web pages connected through the Internet hyperlinks, strings of text that enable surfing the WWW, whose nature is particularly suitable for this metric analysis.

Although the Internet hyperlinks may refer to different functions (i.e., authorizing, commenting, exemplifying, etc.) (Harrison, 2002), the essential feature for Webometrics procedures is their directionality.

Indeed, Internet hyperlinks are directional, pointing from a page to another one; hence, it is possible to distinguish between the "outgoing" links (i.e., number of hyperlinks pointing to other Web pages importing digital information) and the "incoming" links (i.e., number of links received from other Web pages exporting digital information) (Cioleck, 2002). Second, because these hyperlinks are included into a Web page or site characterized by a domain name, it is easy to assign (under the above mentioned limitations) the ability to offer or demand digital information and contents to a particular player (i.e., country, region, institution or organization). Thus, Internet hyperlinks allow analysts to study the relational structure of the WWW (Cioleck, 2002; Han Woo Park, 2003; Maggioni & Uberti, 2003; Uberti, 2004).

Hence hyperlink based indicators capture the relevance of a Web page or site according to its references (outgoing links) or citations (incoming links) (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2001; Rousseau, 1997; Thelwall & Smith, 2002). An example of an index calculated in Webometrics analyses is the "Web impact factor," a measure similar to the impact factor calculated in bibliometrics that captures the influence of a site across the whole Web, calculating the number of "sitations," or incoming links, from other sites (Almind & Ingwersen, 1997; Smith, 1999).

Some critics highlight possible drawbacks related to the use of hyperlinks as useful indicators. The first critic refers to the fact that, since inserting an Internet hyperlink in a Web page is a simple and relatively inexpensive action, the informational content of such an indicator is low. The second relates to the fact that different categories of Web sites (i.e., commercial, institutional, academic) may use hyperlinks in totally different ways.

The answer to the first critic highlights that – since the physical space in a computer screen and, above all, the surfer's attention, are limited – there is a "non-monetary" budget constraint that acts as a powerful disciplining mechanism in forcing the Web