Measuring and Mapping the World Wide Web through Web Hyperlinks

Mario A. Maggioni
Università Cattolica, Italy

Mike Thelwall
University of Wolverhampton, UK

Teodora Erika Uberti
Università Cattolica, Italy

THE WEB: A COMPLEX NETWORK

The Internet is one of the newest and most powerful media that enables the transmission of digital information and communication across the world, although there is still a digital divide between and within countries for its availability, access, and use. To a certain extent, the level and rate of Web diffusion reflects its nature as a complex structure subject to positive network externalities and to an exponential number of potential interactions among individuals using the Internet. In addition, the Web is a network that evolves dynamically over time, and hence it is important to define its nature, its main characteristics, and its potential.

THE INTERNET AND THE WEB

In order to investigate the nature of the Web, it is essential to distinguish between the physical infrastructure (which we will call the "Internet") and the World Wide Web (mostly known as *www*). The Internet is a series of connected networks, each of which is composed of a set of Internet hosts and computers connected via cables, satellites, and so forth. The Web is hosted by the Internet with e-mail as another popular service. The Web is a collection of Web pages and Web sites, many interconnected through 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, a region, or a city) to enable effective and efficient exchanges of digital information; while the Web is a virtual space reflecting the ability to create and export digital information. Of course, the latter would not exist without the former (Berners-Lee & Fischetti, 1999).

While the Internet has a relatively stable infrastructure (because the investments to implement and maintain it are large and costly and few key organizations are involved: mostly corporate, governmental, or nongovernmental organizations), the Web changes very rapidly over time (because it is relatively cheap and easy to create and maintain a Web site and the number of people involved is huge). It is therefore difficult to give a precise and up-to-date description of the Web. For technical reasons, it also defies precise description (Thelwall, 2002).

The most common indicator of Internet diffusion is the number of Internet domain names, which should indicate the ability of a given geographical area to create digital contents and support the exchange flows of information. Unfortunately, this concept is ambiguous and its measurement does not entirely capture the actual diffusion of the Web. First, most generic top level domains (gTLDs), which accounted for almost 67% of the total domains in January 2004 and 56% in January 2007, do not reflect any specific geographical location. Second, some country code top level domains (ccTLDs), even if nominally geo-located, display a mismatch between the official location of the TLD and the actual source of digital information. For example, the .tv domain (an acronym for Tuvalu Islands) is very diffused among television companies internationally because of its acronym, and hence most .tv Web sites are not related to the owning country. Similarly, .nu (an acronym for Niue Islands) is quite common among commercial sites playing on the phonetic similarity between "*nu*" and "*new*", but not necessarily because Niue inhabitants create digital contents for the Web. Third,

even if considered jointly with other technological and economic indicators (e.g., the number of computers or telephone lines), the number of Internet domains may capture a large share of the Internet infrastructure, they do not reveal digital information flows.

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

MEASUREMENTS FROM SEARCH ENGINES

Many Webometric studies need to count the number of Web pages or hyperlinks in one or more Web sites. Although there is special purpose Web crawler software that can do this, such as SocSciBot (socscibot.wlv.ac.uk), most researchers use special commands in commercial search engines. To estimate the number of pages in a Web site, the advanced search site can be used in Google, Yahoo!, or Live Search. For example, to estimate the number of pages in the BBC Web site, the search `site:bbc.co.uk` could be entered as a search in one of the three search engines. Similarly, to count the number of pages linking to the BBC Web site, the search `linkdomain:bbc.co.uk` could be submitted to Yahoo! (only). The two advanced search commands site: and linkdomain: are hence very powerful and have been used in many Webometric investigations (Thelwall, Vaughan, & Bjorneborn, 2005), for example, to compare the sizes of international collections of Web sites and to analyse structures of hyperlinking between them.

One drawback with using commercial search engines for research data is that they do not crawl the whole Web, and perhaps less than 16% of the static pages (Lawrence & Giles, 1999). This means that their results are always likely to be underestimates. Moreover, search engine results can vary unexpectedly over time (Rousseau, 1999) and hence it is difficult to know how reliable their results are, even for comparison purposes. Finally, search engines can estimate different total numbers of results on different results pages (Thelwall, 2008). For example, the first page may report that the estimated number of matches to the query is 45,000, whereas the second page may estimate only 30,000. Nevertheless, recommendations have been made for dealing with this situation and for understanding the results (Thelwall, 2008).

LINK ANALYSIS

A relevant problem in the analysis of the Web concerns measurement. Almind and Ingwersen (1997) adopted quantitative techniques, derived from bibliometric and infometric procedures to analyse the structure and the use of information resources available on the Web. Hence, they introduced the term Webometrics: the bibliometric study of Web pages.

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

The starting point of Webometrics takes into account the structure of the Web: a network of Web pages connected through the Internet hyperlinks, strings of text that support Web navigation, which are particularly suitable for this metric analysis. Although hyperlinks may perform different functions (e.g., authorising, commenting, and exemplifying) (Bar-Ilan, 2004; Harrison, 2002), the essential feature for Webometrics procedures is their directionality.

Since hyperlinks are directional, it is possible to distinguish between the "outgoing" links (i.e., hyperlinks pointing to other Web pages "importing" digital information) and "incoming" links (i.e., links received from other Web pages "exporting" digital information). Second, because these hyperlinks are included in a Web page or site characterised by a domain name, it is easy to assign (subject to the above mentioned limitations) the ability to offer or request digital information and contents to a particular player (i.e., country, region, institution or organisation). Thus, hyperlinks allow analysts to study the relational structure of the Web (Park, 2003; Thelwall, 2005).