In D. Buhalis, W. Höpken & U. Gretzel (Eds.), Information and Communication Technologies in Tourism 2009 - Proceedings of the International Conference in Amsterdam (NL). Wien: Springer, pp. 309-318

On the Importance of Hyperlinks: a Network Science Approach

Rodolfo Baggio[a],
Magda Antonioli Corigliano[a]

[a] Master in Economics and Tourism, Bocconi University, Milan, Italy
{rodolfo.baggio, magda.antonioli}@unibocconi.it

Abstract

Hyperlinks are the essence of the World Wide Web. Their importance is very high due to their ability to provide a visitor with a wealth of good quality information and for the role they play in the ranking of sites by modern search engines. This paper provides a network science approach to provide evidence to the importance of hyperlinking. We examine the webgraph of a tourism destination using graph theoretic methods to highlight the effects that the topological structure has on its navigability. Moreover, through a series of simulations performed on the representation of the real web network we show how a modest increase in the number of links may improve the visibility and the navigability of the destination's webspace.

Keywords: Web navigation, hyperlinks, complex networks, random walks.

1 Introduction

Hyperlinks are the essence of the World Wide Web (WWW). They provide rapid access to segmented information chunks in non-sequential order, mimicking the associative non-linear process used by an individual looking for information (Conklin, 1987). The links between materials on one site and references provided by other sites on the WWW is regarded as a core characteristic of communication (Benkler, 2006). Hyperlinks were a main design feature of the WWW and one of the objectives of the original scheme proposed by Berners-Lee, who wrote that "a 'web' of notes with links (like references) between them is far more useful than a fixed hierarchical system" (1989: 4).

Hyperlinks have further increased their importance due to the extensive commercialisation of the Web. They may provide visitors to a website with access to a wealth of related information which, if well chosen, can be of high value and strengthen their appreciation of that particular site. On the other hand, the mechanisms underlying their creation favour the build up of 'communities of

interest', i.e. groups of websites related by a common topic. This eases the movement between them (surfing) and provides added value to the visitors (Park & Thelwall, 2003; Vaughan et al., 2006). The nature of the links between sites is one of the major determinants of a website's positioning on the results pages of modern search engines and thus vital to online success (Biever, 2004). Hyperlinks also provide a representation of social links between the individuals who own the websites (Adamic & Adar, 2001, 2003; Park, 2003). In this manner, hyperlinks have acquired an economic value, becoming what may be called the *currency* of the Web (Walker, 2002).

Nonetheless, a great number of websites (and particularly in the tourism field) exhibit very few external links, seeming to suffer form a marked *linkphobia*. This low propensity to reference the external world can be ascribed to the 'transfer' to the Web of a historically independent tradition in the conduct of the (mostly) small and medium enterprises forming European tourism (Bramwell & Lane, 2000; Leidner, 2004). Tourism stakeholders appear not to realise how detrimental this can be to the user experience, and the problems this may create in the overall visibility of their websites on the WWW (Baggio, 2006). The aim of this paper is to demonstrate the importance of hyperlinks using a network science approach based on graph theoretic methods. The remainder of this paper is organised as follows. Section 2 presents the background for this research; Section 3 examines the methods used. Results and discussion are reported in Section 4.

2 Background

The study of the Web as a complex network of interconnected elements (the websites bound by the hyperlinks connecting them) has shown that, far from exhibiting some kind of regular or random distribution of connections, it has a well definite structure. The well known *bow-tie* model (Broder et al., 2000; Dill et al., 2002; Flake et al., 2002) sees the Web as a self organizing, self-similar structure, basically divided into the following components (see Fig. 1):

- a core of strongly connected nodes (SCC), in which web pages are joined with bidirectional links;
- a set of pages (IN) connected in a unidirectional way to the SCC (pages have outgoing links that reach the SCC, but are virtually unreachable by other parts of the web);
- a set of pages (OUT) reachable by the those in SCC, but whose links are mainly inward bound (i.e. there are paths from SCC to OUT, but there is no direct connection from OUT to SCC or IN)

- some TENDRILS: pages connected to either IN or OUT without linking to SCC; and
- TUBEs: direct connections between IN and OUT without passing through SCC.

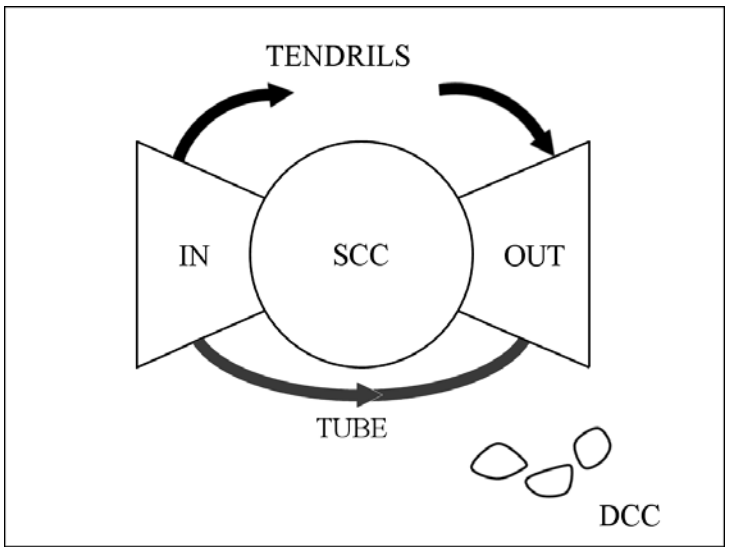

Fig. 1. The bow-tie model of the Web

This structure, along with the other graph theoretic characteristics of the WWW network (distributions of connections, clustering or modularity etc.), is also found in subsections of the Web (self-similarity, see Dill et al., 2002). This topology is the basis for a number of functional algorithms for crawling (Deo & Gupta, 2001; Skopal et al., 2003), searching and communities discovery (Gibson et al., 1998; Newman & Girvan, 2004). These algorithms are fundamental components of the next generation of crawlers, spiders or other automated Web searching tools, and are based on the possibility of identifying well connected groups (communities) of websites sharing common content akin to the one expressed in the query terms user by a web-surfer. Due to its high density of connections, the bow-tie's SCC is a preferred starting point for these explorations. Areas of the WWW with topologies significantly different will inevitably thwart these crawling strategies (Adamic et al., 2001; Deo & Gupta, 2001; Kleinberg, 2006).

Locating a website of interest among the billions present online can be a demanding task for an unaided web-surfer and therefore a large proportion of users rely on a

search engine, with the most using Google (statistics on search engine usage are published by Search Engine Watch, http://searchenginewatch.com/). Consequently, any organisation seeking 'visibility' strives to obtain a good position on the search engine results list. Once a 'starting point website' is found it is quite common that the users continue their navigation guided by the hyperlinks they find on that site (Pan & Fesenmaier, 2006). Movement between linked sites accounts for a large proportion of the visits to websites.

A good indicator of the probability of finding a website is provided by the PageRank. This metric and the underlying algorithm that calculates it were devised by Larry Page and Sergey Brin at Stanford University (Brin & Page, 1998; Page et al., 1999) and form the foundation of the success attained by Google (Vise & Malseed, 2005). PageRank assigns a measure of relevance or importance to each web page, allowing Google to return high-significance pages in response to a user query. The recursive nature of the algorithm, where a page is highly ranked if it is linked to by other highly ranked pages ensures good robustness and reliability. The details of this algorithm are provided in Berkhin (,2005) and Langville and Meyer (2005, 2006). The PageRanking process may be interpreted stochastically as a random walk (Langville & Meyer, 2005; Page et al., 1999). Assume that a user is currently browsing a certain page. After having read it, the surfer moves with a certain probability p to a page that is linked to that page. If there are no links to follow, the user selects (with probability $1-p$) a new page, chosen uniformly over all other Web pages. The PageRank (when normalised) corresponds to the invariant measure for this process. In other words, it represents the long-run proportion of visits made to the destination page, i.e. the probability of a visit to that page.

Let us now suppose a tourist looking for a place to spend his holidays has found a website belonging to a tourism organisation which is connected to a destination. From the 'destination' point of view, it is important to make any effort to 'retain' the visitor for the longest possible time on websites which deal with and therefore belong to the destination. In this way the probability of the destination being chosen as a place for a visit is greatly enhanced. Thus we may see that the structure of the destination's websites network is a crucial element in a tourist's information search process.

In the following sections we analyse a case and show how the density of linkages affects the 'dwell time in destination's pages' of a user surfing websites presented online by the stakeholders of a tourism destination.

3 Data and Methods

The destination used as a test case for this research is the island of Elba, Italy. Elba is a well known seaside destination, off the coast of Tuscany, Italy in the Tyrrhenian Sea. The structural characteristics of the websites belonging to tourism operators located on the island have been discussed elsewhere (Baggio, 2007; Baggio et al., 2007). Here those results can be summarised by saying that the general topology of the network is similar to the one characterising many other complex networked systems (Pastor-Satorras & Vespignani, 2004).

In particular the Elban network exhibits a clear scale-free (power-law) degree distribution (a few nodes have many connections, while the majority have very low connectivity), very poor connectivity and limited modularity. In addition, the structure of the Elban web graph is markedly different from the bow-tie shape found in the Web (Baggio, 2007; Baggio et al., 2007). The network also shows a limited connectivity with regards to the rest of the WWW. The average number of links on a site to websites not belonging to the destination is 1.56 and almost 43% of the websites have none. Moreover, the in- and out- degree distributions show a power law with an exponent much lower than expected from the literature (Pastor-Satorras & Vespignani, 2004). This skewed and 'sparse' distribution of web links provides a very low propensity to reference the external world (Baggio, 2007).

With the existing Elban network as a basis we may then simulate the behaviour of an Internet user looking for information about the destination. This is done here by exploring the system with a series of random walks. The method is far from new in graph theory and network science. Random walks have been used as a way of describing the static and dynamic characteristics of complex networks (da Fontoura Costa et al., 2007 ; Dall'Asta et al., 2005; Yang, 2005), but also as a technique to identify modular communities (Latapy & Pons, 2006) or to measure the 'quality' of a Web search engine (Henzinger et al., 1999).

Let us assume a user has found a website belonging to a tourism operator located on the Elba island (one node of the network is chosen at random). The 'agent' performs a random walk by following the hyperlinks found on the website(s) visited. In other words, one of the links present is chosen with uniform probability and followed to the next website. Here the process is repeated until a website with no links is found. In this case we assume the user changes 'area' and leaves the destination. A maximum of 10 'hops' are allowed and no website can be visited twice (technically this is called a self-avoiding random walk). The algorithm is then run generating 1000 random walks and the average length and the proportion of zero-length walks measured. Further, to show the importance of increasing the general connectivity of the network,

this process was repeated 10 times with different parameters. Each successive simulation was created by adding 2% of links to the previous case. Links were added following a simplified preferential attachment rule (Albert & Barabási, 2002). They were created by randomly choosing a node among the 50% having lowest degrees and connecting it to one chosen (randomly) among the 50% with the highest degrees. This procedure preserves the basic topological characteristics of the network.

4 Results and discussion

The Elban Web network is depicted in Fig. 2. This drawing has been obtained by applying a Fruchterman-Rheingold visualisation algorithm (Fruchterman & Rheingold, 1991) which is especially suited to highlight the disconnected components of the network (this picture is drawn with Pajek, a large network visualisation and analysis program written by Batagelj & Mrvar, 2007).

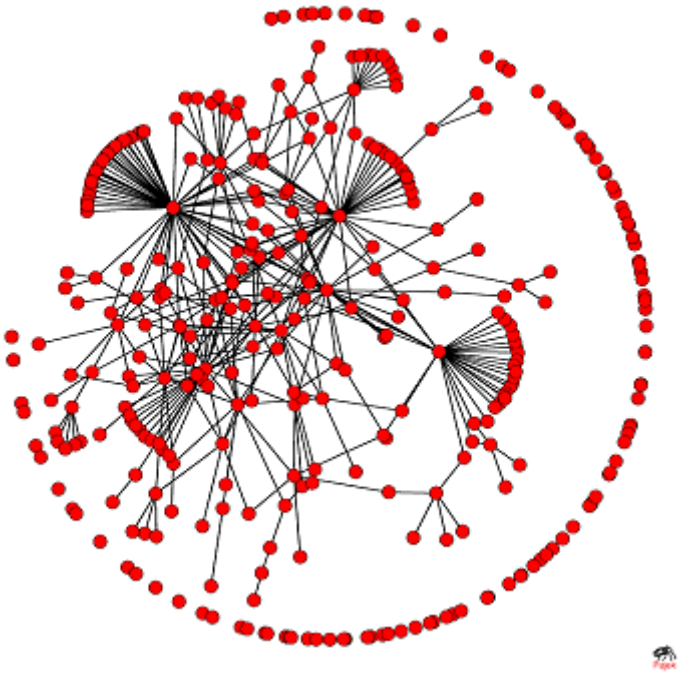

Fig. 2. The webgraph of tourism operators at Elba

A confirmation of this poor connectivity is provided by considering the PageRanks of the Elban tourism websites. Their PageRank distribution was obtained by querying Google and is shown in Fig. 3. The average value is 2.85±0.11 (on a 0 [poor] to 10 [good] scale). This is a low value indicating poor visibility for these websites on the Net.

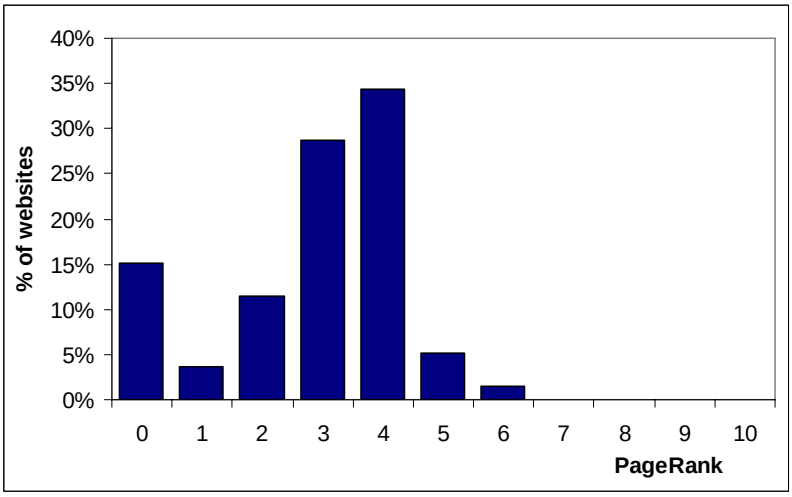

Fig. 3. PageRank distribution for the Elban tourism websites

The analysis of the random walks conducted over this network shows interesting characteristics. The results are shown in Table 1. The first column contains the percentage of links added with respect to the first entry (the original Elban network). Other data shown are: density of links in the network (Density), average path length (Length) and diameter (Diameter), the average length of the random walks (aveRW) and fraction of walks with zero length (%zeroRW).

As described in Section 3 these values were obtained averaging over 1000 random walks. The random walk measurements give an indication of the 'deepness' (and could be related to the time spent online) of the visit to the destination's websites, while density, average path length and diameter are standard measurements of a network's topological characteristics and give indication of its compactness and the ease of navigating it (da Fontoura Costa et al., 2007).

Table 1. Results of the random walk simulations on Elba tourism networks (see text for a description)

Added Links	Density	Length	Diameter	aveRW	%zeroRW
Original	***0.0045***	***3.7***	***10.0***	***0.8015***	***72.4%***
2%	0.0047	3.7	10.0	0.8411	70.6%
4%	0.0052	3.8	9.3	0.9872	65.8%
6%	0.0060	3.7	8.7	1.1980	59.7%
8%	0.0075	3.7	7.6	1.5426	51.5%
10%	0.0097	3.5	6.7	1.9678	43.4%
12%	0.0134	3.2	5.8	2.4937	35.7%
14%	0.0196	2.9	5.0	3.3864	24.5%
16%	0.0301	2.6	4.0	4.7269	14.0%
18%	0.0487	2.3	3.3	6.1465	8.1%
20%	0.0828	2.0	3.0	7.2291	4.5%

The original network, which as discussed above is characterised by a very low connectivity, has a very short random-walk length and a high proportion of zero-length walks. This is also evident when comparing (Table 1 and Fig. 3) these values with the average path length (the average length of the shortest paths between any two nodes) and the diameter (the maximal shortest path).

By increasing the link density, rather obviously, aveRW increases and the proportion of zero-length walks reduces. As long the average random walk as is lower than the network diameter (or average path length) the random walker will have a limited knowledge of the whole network, and his browsing sessions will be much shorter than the topology of the network could allow.

We can describe the results in the following way. After having found a website belonging to Elba, in approximately 73% of the cases a Web surfer changes destination (no more links are found). In the remaining 27% of the cases our visitor is likely to visit on average less than one other website. This is a poor outcome for a tourism destination seeking to compete to attract tourists and to convince them of the worth of a stay. A single website might also be sufficient to provide all the information sought after, but, unfortunately, Elban tourism websites do not perform well from this point of view (see e.g. Tallinucci & Testa, 2006). The situation is even worse when taking into account (Sections 2 and 3) the structure of the Web and likely future developments of search engines where websites with low PageRank will be

difficult to find and sparse networks or small SCC components hinder efficient crawling processes.

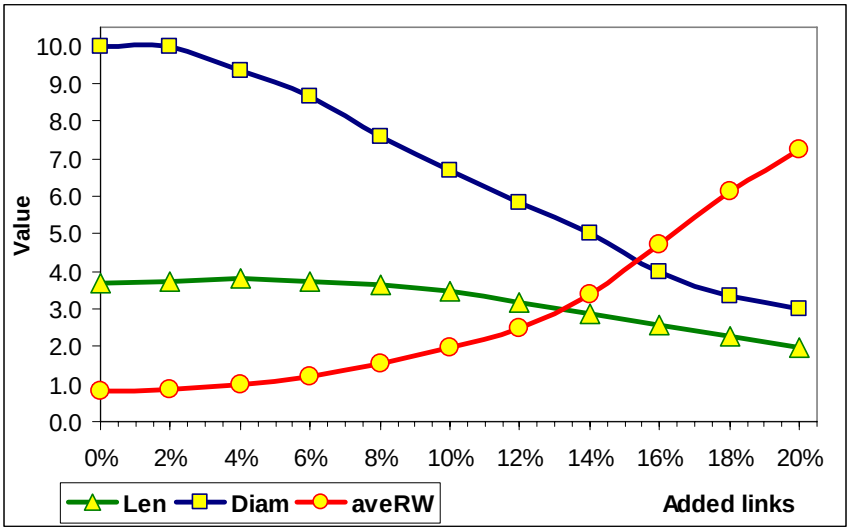

Fig. 3. Modifications of diameter, average path length and average random walk when increasing the number of links

A simple and efficient way to address this situation exists, provided the *linkphobia* of Elban tourism operators can be circumvented. Consideration of Table 1 indicates that the average random walk grows with an increase in link density. This is an expected outcome when increasing the network density. What is interesting to note, however, is that a *moderate* increase in the number of links has the effect of greatly improving the overall navigability of the system. This is mainly due to the topology of the network. This behaviour is more evident when looking at Fig. 3 in which the average random walks for the different augmented networks are shown along with the average path length and the diameter of the same networks. As can be seen, a mere 13% increase in the number of links is sufficient for the random walk to assume the same value of the average path length and a 15.5% increase allows having the diameter's value. Thus modestly improving linking may provide a marked improvement in the capabilities to navigate the whole system.

5 Concluding remarks

Graph theoretic methods (analysis and simulation of random walks) applied to a tourism destination's webspace have allowed us to provide more arguments to the stated importance of the presence of hyperlinks on website's pages. Although limited to a single instance (Elba Island) and using a simplified algorithm, the results presented here appear generally valid and can be extended easily to other cases. Further work is needed to refine and broaden these results and cross-check them with an estimation of the improvement in visibility (e.g. by estimating possible variations in PageRank values).

The outcome and the method used in this study have a strategic relevance for destination managers as well as for individual tourism operators. They provide a way of assessing their own specific context and provide them with simple means of improving the visibility of their websites on the major search engines. Collaboration and cooperation are long discussed arguments in tourism destinations' studies. Here we have provided tangible reasons for doing so in the virtual world, and a very simple, inexpensive and effective way to greatly enhance the possibility of improving web visibility for both a destination and its stakeholders. Adding links connecting more websites within the destination (and to external entities) can be done in a few minutes but can lead to long-term returns.

References

Adamic, L. A., & Adar, E. (2001). You are what you link. *Proceedings of the 10th International World Wide Web Conference, Hong Kong*.

Adamic, L. A., & Adar, E. (2003). Friends and Neighbors on the Web. *Social Networks, 25*(3), 211-230.

Adamic, L. A., Lukose, R. M., Puniyani, A. R., & Huberman, B. A. (2001). Search in Power-Law Networks. *Physical Review E, 64*, 46135-46143.

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Review of Modern Physics, 74*, 47-91.

Baggio, R. (2006). Complex systems, information technologies and tourism: a network point of view. *Information Technology and Tourism, 8*(1), 15-29.

Baggio, R. (2007). The Web Graph of a Tourism System. *Physica A 379*(2), 727-734.

Baggio, R., Antonioli Corigliano, M., & Tallinucci, V. (2007). The websites of a tourism destination: a network analysis. In M. Sigala, L. Mich & J. Murphy (Eds.), *Information and Communication Technologies in Tourism 2007 - Proceedings of the International Conference in Ljubljana, Slovenia* (pp. 279-288). Wien: Springer.

Benkler, Y. (2006). *The Wealth of Networks - How Social Production Transforms Markets and Freedom*. New Haven and London: Yale University Press.

Berkhin, P. (2005). A survey on PageRank computing. *Internet Mathematics, 1*, 73-120.

Berners-Lee, T. (1989). *Information Management: A Proposal*. Geneva, CH: CERN. Retrieved January, 2008, from http://www.w3.org/History/1989/proposal.

Biever, C. (2004). Rival engines finally catch up with Google. *New Scientist, 184*(2474), 23.

Bramwell, B., & Lane, B. (2000). *Tourism Collaboration and Partnerships: Politics Practice and Sustainability*. Clevedon, UK: Channel View Publications.

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual (Web) Search Engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-117.

Broder, A. Z., Kumar, S. R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. L. (2000). Graph structure in the web. *Computer Networks, 33*(1-6), 309-320.

Conklin, J. (1987). Hypertext: An Introduction and Survey. *IEEE Computer, 20*(9), 17-40.

da Fontoura Costa, L., Rodrigues, A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167-242.

da Fontoura Costa, L., Sporns, O., Antiqueira, L., das Graças Volpe Nunes, M., & Oliveira, O. N. J. (2007). Correlations between structure and random walk dynamics in directed complex networks. *Applied Physics Letters, 91*, art.:054107.

Dall'Asta, L., Alvarez-Hamelin, I., Barrat, A., Vázquez, A., & Vespignani, A. (2005). Statistical theory of Internet exploration. *Physical Review E, 71*, art.:036135.

Deo, N., & Gupta, P. (2001). Graph-Theoretic Web Algorithms: An Overview. In T. Böhme & H. Unger (Eds.), *Lecture Notes in Computer Science* (Vol. 2026, pp. 91-102). Berlin: Springer.

Dill, S., Kumar, S. R., McCurley, K., Rajagopalan, S., Sivakumar, D., & Tomkins, A. (2002). Self similarity in the web. *ACM Transactions on Internet Technology (TOIT), 2*(3 - August), 205-223.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2002). Self-Organization of the Web and Identification of Communities. *IEEE Computer, 35*(3), 66-71.

Gibson, D., Kleinberg, J., & Raghavan, P. (1998). Inferring Web communities from link topology. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*, 225-234.

Henzinger, M. R., Heydon, A., Mitzenmacher, M., & Najork, M. (1999). Measuring index quality using random walks on the Web. *Computer Networks, 31*(11), 1291-1303.

Kleinberg, J. M. (2006). Complex networks and decentralized search algorithms. *Proceedings of the International Congress of Mathematicians, Madrid, Spain*.

Langville, A. N., & Meyer, C. D. (2005). Deeper inside PageRank. *Internet Mathematics, 1*, 335-380.

Langville, A. N., & Meyer, C. D. (2006). *Google's PageRank and beyond*. Princeton: Princeton University Press.

Latapy, M., & Pons, P. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications, 10*(2), 191–218.

Leidner, R. (2004). *The European Tourism Industry. A multi-sector with dynamic markets. Structures, developments and importance for Europe's economy*. Luxembourg: Office for Official Publications of the European Communities.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 26113.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the Web* (Working Paper No. SIDL-WP-1999-0120): Stanford Digital Library Project, Stanford University, CA.

Pan, B., & Fesenmaier, D. R. (2006). Online Information Search: Vacation Planning Process. *Annals of Tourism Research*, 33(3), 809-832.

Park, H. W. (2003). Hyperlink Network Analysis: A New Method for the Study of Social Structure on the Web. *Connections 25*(1), 49-61.

Park, H. W., & Thelwall, M. (2003). Hyperlink Analyses of the World Wide Web: A Review. *Journal of Computer Mediated Communication [On-line], 8*(4). Retrieved March 2006, from http://jcmc.indiana.edu/vol8/issue4/park.html.

Pastor-Satorras, R., & Vespignani, A. (2004). *Evolution and structure of the Internet - A Statistical Physics Approach*. Cambridge, UK: Cambridge University Press.

Skopal, T., Snášel, V., Svátek, V., & Krátký, M. (2003). Searching the Internet Using Topological Analysis of Web Pages. *Proceedings of the 2003 International Conference on Communications in Computing (CIC'03), Las Vegas, NV*, 271-277.

Tallinucci, V., & Testa, M. (2006). *Marketing per le isole*. Milano Franco Angeli.

Vaughan, L., Gao, Y., & Kipp, M. (2006). Why are hyperlinks to business Websites created? A content analysis. *Scientometrics, 67*(2), 291-300.

Vise, D., & Malseed, M. (2005). *The Google Story: Inside the Hottest Media and Technology Business of Our Time*. New York: Delacorte Press.

Walker, J. (2002). Links and power: the political economy of linking on the Web. *Proceedings of the 2002 ACM Hypertext Conference, Baltimore, MD*, 72-73.

Yang, S.-Y. (2005). Exploring complex networks by walking on them. *Physical Review E, 71*, art.:016107.