A study of global and local visibility as web indicators of research production

M. R. Martínez-Torres* and M. C. Díaz-Fernández

*Department of Management and Marketing, Facultad de Turismo y Finanzas, University of Seville,
San Francisco Javier, s/n 41018 Sevilla (Spain)
Corresponding author. E-mail: rmtorres@us.es

The concept of webpage visibility is usually linked to search engine optimization, and it is based on global in-link metric, that is, the number of received links from other websites, but without considering the sources of these links. The purpose of this article is to demonstrate that this global idea of visibility is only weakly correlated with web metrics measured over a network of related institutions or organizations (local visibility) and research production. As a case study, global and local visibility measurements have been obtained for a set of Spanish Universities, and they have been correlated with results provided by international rankings like the Webometrics Ranking of World's Universities and the Academic Ranking of World Universities by Shanghai Jiao Tung University. Obtained results suggest that the development of web indicators to be included as part of Universities evaluation programs should consider a local idea of visibility, considering a certain geographical context or similar related institutions.

1. Introduction

Evaluation criteria of research and academic activities are still based on traditional schemes, which ignore informal channels for dissemination of results, references not expressed as bibliographic citations or the impact of the commercial search engines (SEs). Typically, visibility is measured in terms of the number of received links from other webpages, although it has been questioned to what extent this value is related to relevance or data quality (Caro et al. 2011). As an alternative, Kiefer and Stein (2005) introduced the idea of domain-dependent visibility. It is defined as the hit count of an SE when searching for the string of a topic in only that domain. It is not only relevant the number of in-links but also the position of a particular website in a certain context. However, they used the domain-dependent visibility on the sense of web domains that can be obtained using search parameters like 'site:', which can discriminate web pages belonging to a specific country or type of organization.

In this article, we extend the idea of domain-dependent visibility to a network of similar or related institutions, and we refer to this idea of visibility as local visibility, in contrast to the global visibility provided by SEs. This article continues the work previously developed by Martínez-Torres et al. (2011), where the main profiles of websites from a structural point of view were extracted by modelling them as graphs and considering several Social Network Analysis (SNA) features. More specifically, 80 Spanish University websites were analysed considering their internal structure and their relationships with other domains as well. In this article, these relationships with other domains are used to obtain some conclusions about the visibility of Spanish University websites, but limited to an academic context. This concept of local visibility is first compared with the traditional idea of global visibility to highlight the idea that, in certain contexts, it is important to consider to which other websites your specific domain is visible. This is especially important in academic contexts where productivity has been demonstrated to be correlated with some web metrics (Thelwall & Harries 2003). This study also demonstrates that the local idea of visibility is a better web indicator of research production than global visibility. By modelling a set of academic websites as a social network, several measurements for

local visibility based on topological features of the resulting network are proposed.

2. Related work

2.1 International rankings of universities

Today, the three most well-known ranking systems are the Performance Ranking of Scientific Papers for World University directed by Higher Education Evaluation and Accreditation Council of Taiwan (the HEEACT ranking, 2007–present), the Academic Ranking of World Universities by Shanghai Jiao Tung University (the ARWU ranking, 2003–present) and the QS World University Rankings by Times Higher Education, (the THE-QS ranking, 2004–9), which split into two independent ranking programs in 2010 (QS World University Rankings and THE World University Rankings). Although the three programs vary in their methodologies, all of them heavily rely on the research production of Universities, above all, scientific papers indexed in the ISI citation index databases (Huang 2011). Only the third one, THE-QS, relies also on peer review and tends to favour the famous and historically established universities. However, none of these programs consider alternative channels like the web.

The web represents a powerful and cheaper communication tool, which is not still being fully exploited by academic institutions (Barrero et al. 2008). Universities deal with a young, literate, and innovative audience, accustomed to using Internet services. Thus, the Internet has become a common means of providing information and marketing services to prospective students (Poock 2006). Aguillo (2009) argued that evaluation of academic institutions should include a new generation of web indicators, which can be considered as proxies in the valuation of the university global performance, taking into account its activities and outputs and their relevance and impact. The most relevant project in this sense is the Ranking Web or Webometrics performed by the Cybermetrics Lab (Spanish National Research Council, CSIC) for providing information about the performance of universities from all over the world based on their web presence and impact (Aguillo et al. 2006). Webometrics uses link analysis for visibility evaluation, as it is a far more powerful tool than citation analysis or global surveys. Visibility is usually measured as the number of received links from other websites. This idea of visibility is closely related to SE optimization in the sense that the more links are received by a website, the more likely it will be indexed by an SE (Vaughan and Thelwall 2004; Wouters et al. 2006; Vigo and Brajnik 2011). Zhang and Dimitroff (2005) discussed how included metadata may affect visibility of web pages, concluding with some final recommendations about metadata implementation.

2.2 SEs

In the case of Google SE, the far most used SE worldwide, its superior quality of search results compared with other SE is based on the popular PageRank algorithm, a sophisticated method to rank web documents (Brin and Page 1998). PageRank of a given page is not only based on the total number of inbound links but also recursively defined by the PageRanks of those pages, which link to this target page (Guo and Zhang 2008). In general, SEs use the concept of visibility based on the assumption that a high number of received links is related to relevance or data quality. However, recent studies have questioned the relationship between data-quality level and visibility. Caro et al. (2011) conducted a study evaluating the data quality in a group of 88 web portals in three domains and compared these results with their position in a visibility ranking. Their study reveals that websites with the best data quality value are not necessarily the most visible ones. Moreover, they claim that SEresults do not provide details regarding data quality, despite of the importance of these results for decision making. Consequently, the idea of visibility is not covering properly data quality or relevance as it was supposed.

The main problem with SEs is that web designers can artificially improve the likelihood that a particular website appears in prominent positions in response to queries by taking specific actions to promote it (Gori and Witten 2005). For instance, PageRank score can be artificially incremented by creating a keyword page that links to a set of artificial pages containing the selected keywords, which in turn link to the target website. This technique is commonly called a 'link farm'.' In addition, words and phrases are inserted specifically to promote visibility. These words or phrases are not necessarily visible to users. Web designers use metadata that does not form part of the visible content or even hidden text, that is, textual content, which your visitors cannot see, but which is still readable by the SEs. A particularly insidious technique is known as 'cloaking'. In this case, website is designed to give a different page depending on whether the page is being requested by a human visitor or an SE. To avoid these problems, we propose a different approach of visibility considering only the network of related organizations. For this purpose, a set of academic institutions from the same country has been chosen as a case study. A specific crawler has been developed to extract the links among them. The complete list of academic institutions (and other institutions referenced from these websites) has been modelled as a graph, defining nodes as institutions and arcs as the links between them. Local visibility is then obtained using different features of the extracted network. These measurements of visibility are compared with the global visibility, considering the whole web. Findings highlight that it is not only relevant the count of links received by an institution but also the position occupied by this institution among the rest of them.

2.3 Academic websites

Performance of academic websites has been frequently considered in webometric studies. The main reason is that universities are stable institutions that have been visible through the web since a long time ago. In fact, some of them were visible from the very beginning of Internet. Besides, via the Internet, universities can disseminate information about its range of services and broaden its pool of potential clients (Vilnai-Yavetz & Tifferet 2012). On the other hand, the content of academic websites has also been linked to their activity. These studies rely on the idea that web data are indicator of scientific activity. Consequently, a set of derived web indicators can be related to academic results (Thelwall 2002a; Thelwall & Harries 2003, 2004; Smith 2008) or bibliometric indicators (Aguillo et al. 2006). Several studies were specifically focused on building web indicators linked to academic production (Thelwall 2001, 2002a; Scharnhorst & Wouters 2006). Using these indicators, academic productivity (Ingwersen 1998; Smith 1999) or visibility (Polanco et al. 2001; Heimeriks & Van Den Besselaar 2006) can be compared in different countries or areas. For instance, Almind and Ingwersen (1997) developed a study about Nordic academic institutions relationships on the web. They concluded that Danish academic websites exhibit lower visibility than other Nordic countries. Similar results were obtained by Ingwersen (1998) using the so-called Web Impact Factor. This study also shows good visibility results for Norway institutions.

Visualization of Information is also another interesting field of research (Chen 2003). The idea consists of visually illustrating the relationships among a set of universities using several criteria. By applying multivariate statistical analysis, universities can be grouped and plotted together (Vaughan 2006; Polanco et al. 2001). However, other research possibilities are offered by network analysis. The idea consists of modelling the relationships among academic institutions as a network where nodes represent academic websites and arcs represent the hyperlinks among them. Following this scheme, several academic areas can be distinguished. For instance, Heimeriks and Van Den Besselaar (2006) grouped European academic websites in four geographical areas: UK, Scandinavia, Germany, and South Europe institutions. Several other studies were also focused on grouping European Universities in sub-networks attending to different criteria (Ortega and Aguillo 2008; Ortega et al., 2008). Basically, obtained results show that resulting groups of academic institutions typically follow linguistic patterns (Heimeriks et al. 2003; Heimeriks 2005; Vaughan 2006) or geographical patterns (Thelwall 2002b; Thelwall et al. 2003). More specifically, the study from Holmberg and Thelwall (2007) concluded that geographical distances negatively affect the network structure of e-government sites for the case study of Finland.

With respect to the case studies considered in these previous research, they are usually focused on countries, such as Spain (Thelwall & Aguillo 2003; Ortega & Aguillo 2007; Martínez-Torres et al. 2011), Canada (Vaughan & Thelwall 2005; Vaughan 2006), New Zealand (Smith 2001), or wider geographical areas, such as the European Union (Heimeriks and Van Den Besselaar 2006; Ortega & Aguillo 2008; Thelwall & Zuccala 2008) or Scandinavia (Ortega & Aguillo 2008).

Previous literature in this field has in common that data are extracted from Internet. Although Internet is a valuable source of information, not all the extracted information is meaningful in the context of a particular topic or study. Two major problems related to link analysis can be distinguished:

- The first one is associated to the coverage of the web provided by SEs. Extracted data can be biased due to the different coverage of a group of considered websites (Mowshowitz and Kawaguchi1 2005).
- The second problem is that the ranking position provided by SEs within the web is set in terms of inlinks and outlinks to other websites (Langville & Meyer 2006). This value is independent of the content of the web page, which then contributes to relevance. As a result, ranking judgements of websites are affected by non-related websites, and it can be artificially improved.

The methodology proposed in this article is intended to solve both problems. First, a specific crawler is designed and implemented to guarantee that considered websites are covered the same way, and then a set of related websites and their relationships is modelled as a graph to extract local measures of visibility.

3. Methodology and visibility approaches

Fig. 1 shows a scheme of the proposed methodology. First, data are extracted using a specific crawler with a programmable depth of link coverage. Each website is modelled as a star shaped network. The central node is the target domain (a given Spanish University), and the rest of nodes represent all the other domains linked from the considered website. To extract each individual domain network, the crawler has to explore the website content looking for hyperlinks to external websites. Once each individual website has been modelled as a graph, the 80 considered websites are merged together to obtain the global domain network, containing the 80 nodes corresponding to the 80 Spanish Universities and all the external domains linked from these websites. Finally, several features of the global domain network are extracted using SNA. These features represent different approaches to the concept of local visibility.
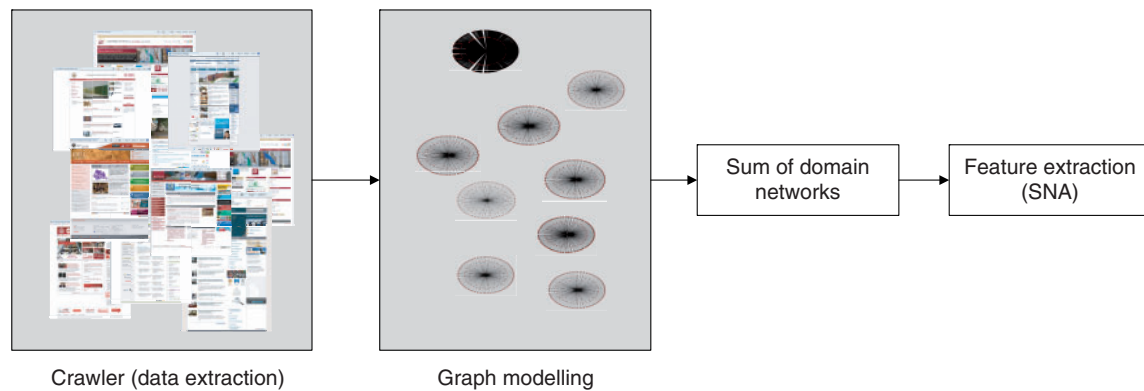
Figure 1. Proposed methodology.

3.1 Data extraction

The availability of a huge volume of data through Internet has lead to an increasing interest in knowledge retrieval from large amount of hypertext (Almpanidis et al. 2007). Two methods for extracting links among websites are available today: crawlers and SEs. A crawler (also referenced as web robot or spider) is an agent that automatically explores the hypertext structure of the web. The crawler starts from a root domain (usually the homepage of a website) or a set of root domains and then continues recursively retrieving all documents referenced by that document until a certain stop criterion is reached. Alternatively, some SEs (i.e. Yahoo!, MSN) can be used for link counting using advanced search commands (Paek et al. 2011). The main limitation of using SEs is that their coverage of websites is sometimes incomplete. Hence, the results can be inaccurate (Thelwall & Zuccala 2008).

In this work, a specific crawler for extracting links among academic websites has been developed. This crawler has been programmed using MATLAB™ (Register 2007). The crawler starts with the root domain, reading its content, processing the content information as strings and finding the links contained in the page. Then it processes the pages found the same way it did with the home page. This process is repeated until the crawler decides to stop attending to a given criterion. Extracted information can be seen as a tree-structure, the root is the start-URL; all links in that root-HTML-page are direct sons of the root, and subsequent links are then sons of the previous sons. An important parameter to be chosen is the page depth of link coverage when capturing website information. A depth of seven has been used in this study. This value is considered sufficient to capture the essential information of website structure, being higher than the depth of five used in some previous studies (Yang & Qin 2008). In fact, some studies point out that just a few levels are necessary to be downloaded by a crawler (no more than 3–5 'clicks' away from the start page) to reach almost 90% of the pages that users actually visit (Baeza-Yates and Castillo 2004).

For the purpose of this study, the main advantage of using a specific crawler is that guarantees an unbiased coverage. The coverage of the major SEs has been proved to be biased, and they cannot guarantee that the set of websites included in this study could be covered to the fixed depth of seven (Vaughan & Thelwall 2004).

3.2 Graph modelling

The developed crawler is able to capture websites structure from a starting root domain and considering a maximum level of seven. The crawler discriminates between pages belonging to the root domain and pages belonging to different domains. As a result, a domain network is generated following the Pajek format (Nooy et al. 2005, Martínez-Torres et al. 2011). The extracted domains are considered as nodes, whereas the arcs represent links from the root domain to those other referenced domains. It is a similar procedure to those used by studies focused on social networks (Shu and Chuang 2011). The resulting domain network is a star-shaped directed graph, with the root domain at the centre of the star and the rest of domains linked with it.

The described process has been repeated for each one of the root domains in a list of academic websites. Finally, the obtained set of domain networks has been merged together to obtain a final network representing the relationship among academic websites in a certain area, as well as the relationship among them and other external institutions.

3.3 Feature extraction

In this article, we make use of web indicators based on SNA and in-link visibility, but measured over a set of interrelated websites. These indicators are based on the following SNA features.

- Degree: it is defined as the number of edges/arcs incident with a node. To make this indicator non-size dependent, it can be normalized by the total number of network nodes. In the case of directed network, it can be distinguished the number of incoming links

(InDegree) and outcoming links (OutDegree) (Toral et al. 2009a). In the context of webometrics, degree can be used as a measure of a web domain visibility (Cothey 2005; Kretschmer and Kretschmer 2006).

- Closeness centralization: it is an index of centrality based on the distance between nodes. The closeness centrality of a node is calculated using the total distance between one node and all other nodes (Mote et al. 2007). This, larger distances, lead to lower closeness centrality scores (Toral et al. 2009b). A high value of closeness centrality means a short average distance from the vertex to any other vertex. Consequently, the vertex will be better positioned in spreading information to other vertices (Okamoto et al. 2008).
- Betweenness: it is an alternative measure of centrality. Instead of using the distance between nodes, betweenness considers a node is more central in accordance with its role as an intermediary in the network (Nooy et al. 2005). The intermediary role means the extent to which this node is performing as a link to facilitate the connection of nodes within the network (Martínez-Torres 2012). A more formal definition of betweenness centrality of a vertex is the proportion of all geodesics between pairs of other vertices in the network including this vertex (being a geodesic, the shortest path between two nodes). In the context of library information systems, betweenness is used to detect gateways or hubs connecting different web networks (Ingwersen 1998; Faba-Pérez et al. 2005).
- K-Neighbour: it refers to the number of neighbours from a particular node located in a distance of k, being the distance defined as the number of ties between two nodes. In directed networks, input and output neighbours can be computed. The concept of neighbours is different to the concept of degree, as the latter takes into account multiple lines among nodes. In the context of webometrics, just 1-input neighbour has been taken into account, as they represent the number of nodes that have directly linked the considered node (Martínez-Torres et al. 2010).
- Brokerage roles: A broker is defined as the central node in a directed triad (being a triad, a set of three connected vertices). The middle node is called coordinator if it is connecting two nodes belonging to the same cluster or partition, and representative if it connects nodes belonging to different partitions, Fig. 2 (Toral et al. 2010a).
- Hubs and authorities: Two types of important vertices can be distinguished in directed networks: hubs and authorities (Kleinberg 1998). In the context of this study, hubs are those websites representing good starting points to find information. As a difference, authorities are those websites that contain useful information on a particular topic (Sarukkai 2000). A vertex

Figure 2. Coordinator and representative developing a brokerage roles.

is said to be a good hub when pointing to many good authorities, and it is said to be a good authority when it is pointed by many good hubs. The notion of 'hubs/authorities' is typically applied on the Web graph structure.

These previous SNA features are measured over the final network resulting from merging together individual academic websites. They represent different approaches to local visibility, as all of them are obtained from the interrelations of a set of websites related to a specific topic.

4. Results

4.1 Case study: Spanish universities network

Eighty Spanish universities have been considered for this study. All of them are included in the Webometrics Ranking of World Universities (www.webometrics.info), where >6000 universities all over the world are sorted according to size and visibility. Table 1 lists the root domains of the considered websites. They cover almost the whole range of Webometrics Ranking and exhibit a variety of sizes in terms of domains and web pages.

Figs 3 and 4 show the domain and page network, respectively, corresponding to the particular case of the University of Seville. In particular, Fig. 3 shows the star-shaped structure of the domain network, with the root domain at the centre of the network connected through arcs to the rest of referenced domains. Each arc has an associated number representing the number of times each external domain has been linked by the considered website (this number is not shown to improve readability of figures). The developed crawler has analysed >718,000 web pages and more than four million outlinks for the 80 considered universities.

Once the 80 domain networks corresponding to the websites listed in Table 1 have been extracted, all of them have been merged together to obtain the Spanish universities network (Fig. 5).

This network (Fig. 5) has a size of 16,957 nodes and includes the 80 Spanish universities considered, as well as the rest of the other domains linked from these websites.

4.2 Visibility results

The resulting network (Fig. 5) has been used to compare the different approaches to universities websites visibility

Table 1. List of considered websites

Spanish university websites

http://www.ucm.es/	http://www.ual.es/
http://www.upc.edu/	http://www.udl.es/
http://www.upm.es/	http://www.ujaen.es/
http://www.uab.es/	http://www.umh.es/
http://www.ehu.es/	http://www.deusto.es/
http://www.ub.edu/	http://www.unavarra.es/
http://www.us.es/	http://www.upct.es/
http://www.upv.es/	http://www.upo.es/
http://www.um.es/	http://www.ie.edu/
http://www.ugr.es/	http://www.upcomillas.es/
http://www.ua.es/	http://www.ceu.es/
http://www.uvigo.es/	http://www.iese.edu/
http://www.uv.es/	http://www.ubu.es/
http://www.uam.es/	http://www.urv.net/
http://www.usal.es/	http://www.unirioja.es/
http://www.uji.es/	http://www.uem.es/
http://www.unizar.es/	http://www.esade.edu/
http://www.usc.es/	http://www.ucam.edu/
http://www.uib.es/ca/	http://www.mondragon.edu/
http://www.uclm.es/	http://www.uvic.es/
http://portal.uned.es/	http://www.cef.es/
http://www.uva.es/	http://www.uch.ceu.es/
http://www.upf.edu/	http://www.nebrija.com/
http://www.unav.es/	http://www.uic.es/
http://www.uc3m.es/	http://www.url.es/
http://www.uniovi.es/	http://www.esdi.es/
http://www.uma.es/	http://www.uax.es/
http://www.uco.es/	http://www.vives.org/
http://www.ull.es/	http://www.uimp.es/
http://www.udc.es/	http://www.ucjc.edu/
http://www.unex.es/	https://www.ucv.es/
http://www.uah.es/	http://www.uspceu.com/
http://www.uoc.edu/	http://www.cesdonbosco.com/
http://www.udg.edu/	http://www.ufv.es/
http://www.ulpgc.es/	http://www.esic.es/
http://www.unican.es/	http://www.cepade.es/
http://www.unileon.es/	http://www.eoi.es/portal/
http://www.urjc.es/	http://www.esmuc.net/
http://www.uca.es/	http://www.udima.es/
http://www.uhu.es/	http://www.eupmt.es/

and to correlate them with international university rankings. Several visibility measurements can be defined in the context of SNA. As a reference, the Webometrics Ranking of World's Universities and the ARWU have been considered.

The Webometrics Ranking is published twice per year (January and July) since 2004, covering >20,000 Higher Education Institutions worldwide (http://www. webometrics.info/). It is based on several indicators, visibility being one of them. In fact, this indicator represents the 50% of the Webometrics rank (Aguillo et al. 2008). The visibility of Webometrics rank has been labelled as V-WEB in our work, and it is defined as the total number of unique external links received (inlinks) by a site, according to Yahoo Site Explorer. V-WEB represents a global visibility measurement, as this value does not

distinguish the source of inlinks. The Shanghai ranking ranks academic institutions on the basis of their research performance, and it has been labelled as ARWU. The rest of indicators of Table 2 derive from the obtained network illustrated in Fig. 5 and represent the different approaches to local visibility used in this work.

Fig. 6 represents the extracted network of 80 Spanish universities, where the size of nodes is proportional to their in-degree value. The most relevant Spanish universities are located in the denser part of this network.

Indicators from Table 2 have been compared using the Spearman's rank-order correlation, which is the non-parametric version of the Pearson product-moment correlation, and measures the strength of the association between ranked variables, that is, how closely several sets of rankings agree with each other. Table 3 details the obtained results.

Basically, Table 3 shows that the global concept of visibility represented by V-Web leads to a ranking of websites different to the one obtained by local approaches of visibility, considering websites belonging to the same academic area. The first column of Table 3 reveals there is a weak correlation among V-Web and the rest of visibility indicators. However, the correlation among local measures of visibility and ARWU ranking is positive and significant. The only exception is BET ranking, which is negatively correlated with the rest of ranking indicators (except V-WEB, where correlation is almost 0, but not significant). This result can be explained because the betweenness centrality is related to websites connectivity, that is, websites that are frequently referenced and which also are referencing a lot of external domains. The rest of local ranking indicators are positively correlated. The reason is that all of them capture the idea of incoming links, considering the number of input neighbours or the number of incident arcs. Coordinators and representatives indicators also take into account incoming and outcoming links. However, they do not consider multiple lines like betweenness indicator.

5. Discussion and implications

The web is an emergent information source which is demanding new research (Spink 2002). More specifically, hyperlinks and scholarly uses of the web as a source of information are of particular interest (Cronin 2001). Although it is truth that web pages are not always easy to count and they are continuously changing or even disappearing over time, several studies confirm a significant correlation between web metrics and other activity or production measurements in academic contexts. New advances about link analysis have been encouraged owing to the success of the PageRank algorithm by Google and other similar link-based approaches adopted by SEs.
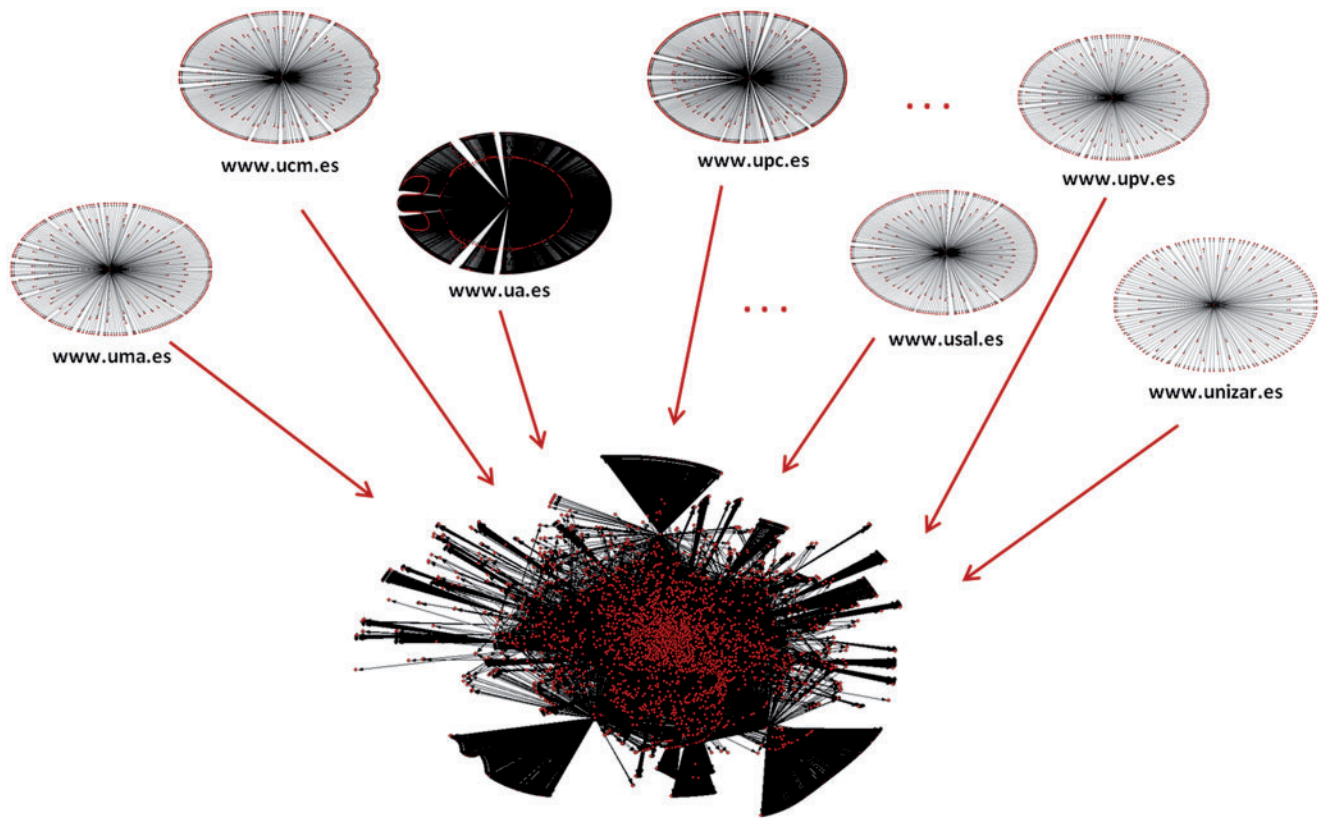
Figure 3. University of Seville domain network.

Figure 4. University of Seville page network.

Some previous research works have dealt with the problem of visibility measurement. For instance, the 'Web Impact Factor' introduced by Ingwersen (1998) is the ratio between the number of web pages in a certain website that receives links from other websites and the number of web pages included in the website. The main problem of this definition is the mathematical artifacts derived from power law distributions of these variables. A new method for visibility evaluation has been

proposed by Espadas et al. (2008). Basically, they consider visibility depends on visitors, SE results and other sites that make recommendations as links. More specifically, they consider the idea of competitor websites included in this method, which is in line with the local visibility approach proposed in this article.

In this work, a specific crawler for extracting links among academic websites has been developed. SEs can also be used for this purpose. However, they exhibit some problems related to their search operator and their databases (Ortega & Aguillo 2009). A network resulting from the aggregation of 80 Spanish universities (all of them are included in the Webometrics Ranking of World Universities) has been considered for this study to compare the different approaches with visibility. As a reference, the visibility of Webometrics rank and the Shanghai rank has been used.

The first conclusion is that the visibility of a university website through the web is not always a good measure of the relative importance of this website in a country or an academic area. More specifically, obtained results show a weak correlation value between V-WEB visibility measurement and the rest of SNA derived indicators, which consider in links coming from an academic context. The main difference is that V-WEB considers all incoming links independently of their origin, whereas the local indicators only consider the small portion of the web including

Figure 5. Result of merging together the 80 Spanish universities domain networks.

Table 2. Considered measures of visibility

Visibility	Indicator	Focus
Visibility according to Webometric rank	V-Web	Global
In-degree	IN-D	Local
Input closeness centrality	IN-CLO	Local
Betweenness centrality	BET	Local
1-Neighbour	1NEIGH	Local
Coordinators	COORD	Local
Representatives	REP	Local
Authorities	AUTH	Local
Shanghai ranking	ARWU	Global

related institutions. V-WEB as a global measure of visibility is enough when working with general information or with institutions interested in being visible for the rest of the world. However, this is not the case of academic institutions where the value of received in-links is not the same depending on their origin.

The second important result is that local visibility is a better measurement of research production, as it is shown by the correlation coefficients with the ARWU index. In academic contexts, linking models focused on universities claim that the pages of 'better' researchers (i.e. with higher research scores) attract more inlinks, as these researchers are able to generate more web pages (Thelwall & Harries 2004). Some refinements of these models consider that better researchers may not only produce more research-related web pages but may also produce more educational web pages. Finally, it has been suggested that that richer universities tend to better fund web-related infrastructure (Thelwall 2004). The position of universities in international rankings is today considered as a strategic issue for most universities. Although they begin as a tool to rank universities by comparing their research capacity and output, they are increasingly being used as a stick to measure institutions and not just in relation with research (Docampo 2011). However, there is a controversy regarding their methodologies and their choice of indicators. Some authors criticize the use 1D measures that can be misleading owing to oversimplification (Dehon et al. 2010). The developed measurement of local visibility offers the possibility of incorporating new web indicators as a part of these rankings.

In general, the main advantage of local visibility is precisely the fact of considering related institutions links, which avoids the possibility of artificially incrementing visibility of websites. The drawback is that local visibility requires a previous knowledge of related institutions to limit the scope of link search.

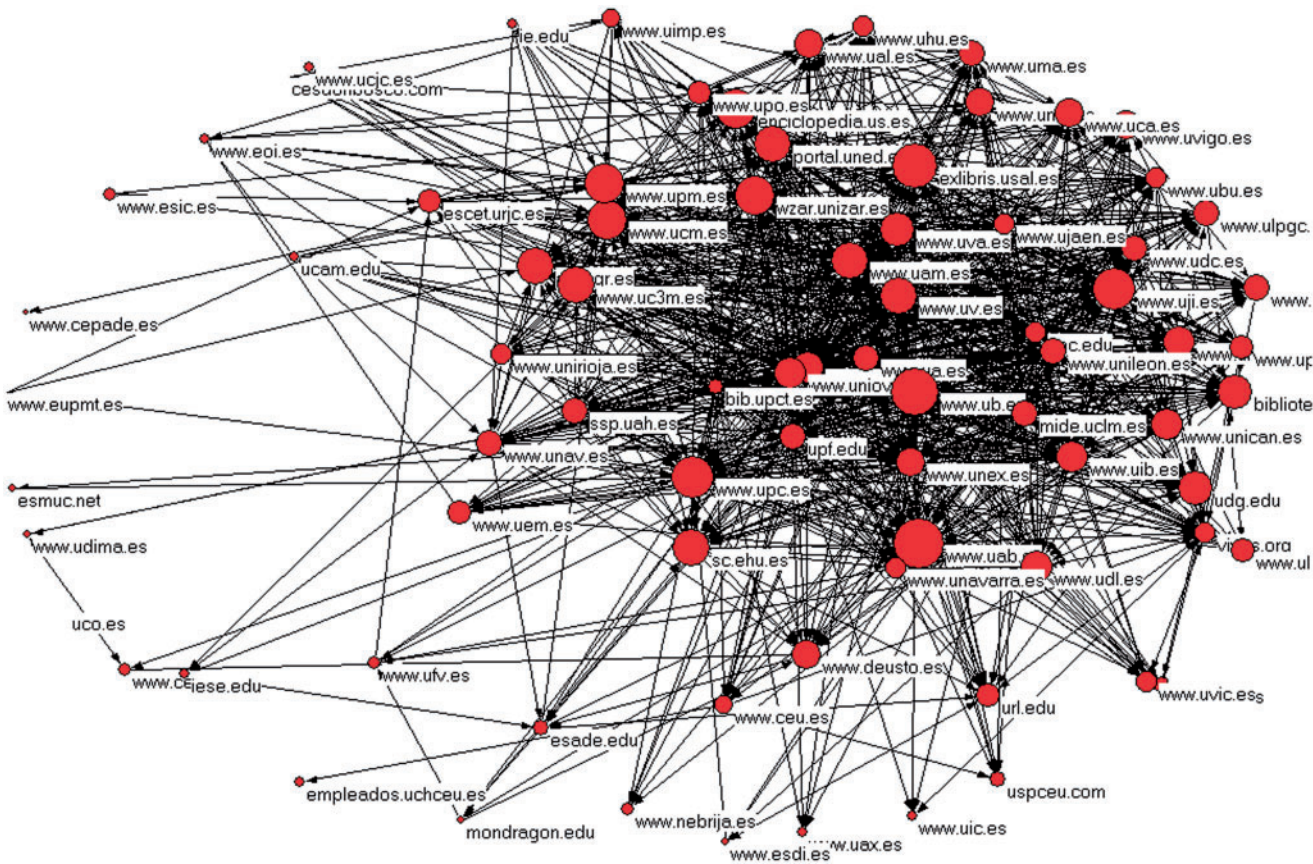One of the consequences of the concept of global visibility is that the majority of websites try to appear in the

Figure 6. Network of Spanish universities.

Table 3. Spearman's rho correlations results

Visibility indicators	V-Web	IN-D	IN-CLO	BET	1NEIGH	COORD	REP	AUTH	ARWU
V-Web	1.000	0.297**	0.310**	−0.090	0.336**	0.304**	0.286**	0.078	−0.196*
IN-D	0.297**	1.000	0.962**	−0.539**	0.956**	0.666**	0.692**	0.710**	0.530**
IN-CLO	0.310**	0.962**	1.000	−0.575**	0.976**	0.708**	0.723**	0.758**	0.586**
BET	−0.090	−0.539**	−0.575**	1.000	−0.571**	−0.323**	−0.370**	−0.382**	−0.344*
1NEIGH	0.336**	0.956**	0.976**	−0.571**	1.000	0.694**	0.716**	0.759**	0.574**
COORD	0.304**	0.666**	0.708**	−0.323**	0.694**	1.000	0.850**	0.631**	0.346*
REP	0.286**	0.692**	0.723**	−0.370**	0.716**	0.850**	1.000	0.583**	0.407**
AUTH	0.078	0.710**	0.758**	−0.382**	0.759**	0.631**	0.583**	1.000	0.311*
ARWU	−0.196*	0.530**	0.586**	−0.344*	0.574**	0.346*	0.407**	0.311*	1.000

*Correlation is significant at the 0.05 level (two-tailed).
**Correlation is significant at the 0.01 level (two-tailed).

first positions on SE pages. Thus, web designers or administrators try to take advantage of SE algorithms to improve the ranking of their website when a user types in relevant keywords in an SE. This trend of promoting visibility is actually forcing SE operators to modify their ranking model in ad hoc and secret ways in an attempt to outwit those who try to exploit it. As an alternative, automatic analysis of page content is gaining central importance in both crawling and page scoring

(Diligenti et al. 2004; Toral et al. 2010b) and can help detect artificial communities conceived solely for page promotion.

The proposed measurement of local visibility is able to avoid some of these problems, as neither artificially created web pages nor metadata would be considered in the network of related institutions. Moreover, local visibility measurement has been shown to be more correlated with rankings exclusively focused on research production, like

the Shanghai index. In fact, the Webometrics Ranking has started to consider as part of the impact sub indicator not only the number of external in links that the university web domain receives from third parties but also their institutional prestige and their academic performance. Therefore, they are extending the idea of link popularity towards link diversity.

6. Conclusion

Website co-link analyses have been used in this article to model relationships among university institutions in Spain and to derive several measurements of local visibility. More specifically, universities and their relationships have been modelled as a graph, and visibility of the different institutions has been obtained considering several features resulting from the analysis of obtained graph from the perspective of SNA. The visibility of Webometrics rank, which is obtained using Yahoo explorer, and the Shanghai index have been used as a reference. Findings reveal that local visibility is weakly correlated with global visibility, which means that the idea of visibility is different if we consider the whole web or if we only consider a smaller portion of the web integrated by related or competitors' websites. This result is explained by the general interest of web designers in promoting the relevance of their web pages, using specific techniques to take advantage of the way in which global visibility is measured. However, and due to the increasing importance of Internet as a data source, it would be useful for decision makers to avoid biased data. This fact is clearly illustrated in academic contexts where specific rankings establish some relations between visibility and academic/research production. Obtained results clearly show that the local measurement of visibility is able to consider the scientific production of universities avoiding biased data by limiting the scope of link search to the specific portion of the web constituted by related institutions. Although this method requires a previous knowledge of related websites, it can benefit in the future from semantic web advances. Among other things, the semantic web makes information more meaningful to people by making it more understandable to machines. As a result, the scope of link search could be limited to a certain topic or area as required by the concept of local visibility.

References

Aguillo, I. (2009) 'Measuring the Institution's Footprint in the Web', *Libray Hi Tech (LHT)*, 27/4: 540–56.

Aguillo, I. F. *et al.* (2006) 'Scientific Research Activity and Communication Measured with Cybermetrics Indicators', *Journal of the American Society for Information Science and Technology*, 57/10: 1296–302.

Aguillo, I. F., Ortega, J. L. and Fernandez, M. (2008) 'Webometrics Ranking of World Universities: Introduction, Methodology and Future Developments', *Higher Education in Europe*, 33/2–3: 233–44.

Almind, T. C. and Ingwersen, P. (1997) 'Informetric Analyses on the World Wide Web: Methodological Approaches to 'webometrics'', *Journal of Documentation*, 53/4: 404–26.

Almpanidis, G., Kotropoulos, C. and Pitas, I. (2007) 'Combining Text and Link Analysis for Focused Crawling—An Application for Vertical Search Engines', *Information Systems*, 32/6: 886–908.

Baeza-Yates, R. and Castillo, C. (2004) 'Crawling the Infinite Web: Five Levels Are Enough', *Lecture Notes in Computer Science*, 3243: 156–67.

Barrero, F., Toral, S. L. and Gallardo, S. (2008) 'eDSPLab: Remote Laboratory for Experiments on DSP Applications', *Internet Research*, 18/1: 79–92.

Brin, S. and Page, L. (1998) 'The Anatomy of a Large-Scale Hypertextual Web Search Engine', *Computer Networks and ISDN Systems*, 30/1–7: 107–17.

Caro, A., Calero, C. and Moraga, M. A. (2011) 'Are Web Visibility and Data Quality Related Concepts?', *IEEE Internet Computing*, 15/2: 43–9.

Chen, C. (2003) *Mapping Scientific Frontiers: The Quest for Knowledge Visualization*. London: Springer.

Cothey, V. (2005) 'Some preliminary results from a link-crawl of the European union research area web'. In: Ingwersen, P. and Larsen, B. (eds) *Proceeding of the 10th International Conference of the International Society for Scientometrics and Informetrics*. Stockholm: Karolinska University Press.

Cronin, B. (2001) 'Bibliometrics and Beyond: Some Thoughts on Web-Based Citation Analysis', *Journal of Information Science*, 27/1: 1–7.

Dehon, C., McCathie, A. and Verardi, V. (2010) 'Uncovering Excellence in Academic Rankings: A Closer Look at the Shanghai Ranking', *Scientometrics*, 83: 515–24.

Diligenti, M., Gori, M. and Maggini, M. (2004) 'A Unified Probabilistic Framework for Web Page Scoring Systems', *IEEE Transactions on Knowledge and Data Engineering*, 16/1: 4–16.

Docampo, D. (2011) 'On Using the Shanghai Ranking to Assess the Research Performance of University Systems', *Scientometrics*, 86/1: 77–92.

Espadas, J., Calero, C. and Piattini, M. (2008) 'Web Site Visibility Evaluation', *Journal of the American Society for Information Science and Technology*, 59/11: 1727–42.

Faba-Pérez, C. *et al.* (2005) 'Comparative Analysis of Webometric Measurements in Thematic Environments', *Journal of the American Society for Information Science and Technology*, 56/8: 779–85.

Gori, M. and Witten, I. (2005) 'The Bubble of Web Visibility', *Communications of the ACM*, 48/3: 115–17.

Guo, C. and Zhang, L. (2008) 'An Improved BA Model Based on the PageRank Algorithm', in *4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM '08*, pp. 1–4. Dalian, China: IEEE Communications Society.

Heimeriks, G. (2005) *Knowledge Production and Communication Research Networks*. Amsterdam: University of Amsterdam.

Heimeriks, G., Horlesberger, M. and Van Den Besselaar, P. (2003) 'Mapping Communication and Collaboration in Heterogeneous Research Networks', *Scientometrics*, 58/2: 391–413.

Heimeriks, G. and Van Den Besselaar, P. (2006) 'Analyzing Hyperlinks Networks: The Meaning of Hyperlink Production', *Cybermetrics*, 10/1 <http://www.cindoc.csic.es/cybermetrics/articles/v10i1p1.html> accessed 9 Jan 2011.

Holmberg, K. and Thelwall, M. (2007) 'Local government web sites in Finland: A geographic and webometric analysis'. In *Proceeding of the 11th International Conference of the International Society for Scientometrics and Informetrics*. Madrid, Spain.

Huang, M. H. (2011) 'Comparison of Three Major Academic Rankings for World Universities: From a Research Evaluation Perspective', *Journal of Library and Information Studies*, 9/1: 1–25.

Ingwersen, P. (1998) 'The Calculation of Web Impact Factors', *Journal of Documentation*, 54/2: 236–43.

Kiefer, P. and Stein, K. (2005) 'Visibility Analysis on the Web as an Indicator for Public Relations and Marketing Evaluation', In *Proceedings of the 2005 International Conference on Computational Intelligence for Modelling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce*, CIMCA-IAWTIC'05, pp .151–7. Vienna, Austria: IEEE Computer Society.

Kleinberg, J. (1998) 'Authoritative Sources in a Hyperlinked Environment', In *Proceedings of the 9th ACMSIAM Symposium on Discrete Algorithms*, pp. 668–77. San Francisco, California, USA: Society for Industrial and Applied Mathematics Philadelphia.

Kretschmer, H. and Kretschmer, T. (2006) 'Application of a New Centrality Measure for Social Network Analysis to Bibliometric and Webometric Data', In *Proceeding of the IEEE international conference on digital information management*, (ICDIM), pp. 199–204. Bangalore, India: IEEE.

Langville, A. and Meyer, C. D. (2006) *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton, NJ: Princeton University Press.

Martínez-Torres, M. R., Toral, S. L., Palacios, B. and Barrero, F. (2011) 'Web Site Structure Mining Using Social Network Analysis', *Internet Research*, 21/2: 104–123.

Martínez-Torres, M. R. (2012) 'A Genetic Search of Patterns of Behaviour in OSS Communities', *Expert Systems with Applications*, 39/18: 13182–92.

Martínez-Torres, M. R., Toral, S. L. and Barrero, F. (2010) 'The Role of Internet in the Development of Future Software Projects', *Internet Research*, 20/1: 72–86.

Mote, J. E. *et al*. (2007) 'New Directions in the Use of Network Analysis in Research and Product Development Evaluation', *Research Evaluation*, 16/3: 191–203.

Mowshowitz, A. and Kawaguchi1, A. (2005) 'Measuring Search Engine Bias', *Information Processing and Management*, 41/5: 1193–205.

Nooy, W., Mrvar, A. and Batagelj, V. (2005) *Exploratory Network Analysis with Pajek*. New York: Cambridge University Press.

Okamoto, K., Chen, W. and Li, X. (2008) 'Ranking of Closeness Centrality for Large-Scale Social Networks', *Computer Science*, 5059: 186–95.

Ortega, J. L. and Aguillo, I. F. (2007) 'La Web Académica Española En El Contexto Del Espacio Europeo de Educación Superior: Estudio Exploratorio', *El Professional De La Información*, 16/5: 417–25.

——. (2008) 'Visualization of the Nordic Academic Web: Link Analysis Using Social Network Tools', *Information Processing and Management*, 44/4: 1624–33.

——. (2009) 'Mapping World-Class Universities on the Web', *Information Processing and Management*, 45/2: 272–9.

Ortega, J. L. *et al*. (2008) 'Maps of the Academic Web in the European Higher Education Area—An Exploration of Visual Web Indicators', *Scientometrics*, 74/2: 295–308.

Paek, H. -J. *et al*. (2011) 'Theories Into Practice: A Content Analysis of Anti-Smoking Websites', *Internet Research*, 21/1: 5–25.

Polanco, X. *et al*. (2001) 'Clustering and Mapping European University Web Sites Sample for Displaying Associations And Visualizing Networks', *NTTS&ETK 2001 Conference*, Hersonissos, Crete.

Poock, M. C. (2006) 'Characteristics of an Effective Web Site in Educational Leadership', *College Student Journal*, 40/4: 785–90.

Register, A. H. (2007) *A Guide to MATLAB Object-Oriented Programming*. Boca Raton, FL, USA: Chapman and Hall/CRC Press.

Sarukkai, R. R. (2000) 'Link Prediction and Path Analysis Using Markov Chains', *Computer Networks*, 33/1–6: 377–86.

Scharnhorst, A. and Wouters, P. (2006) 'Web Indicators—A New Generation of S&T Indicators? *Cybermetrics*', 10/1 <http://www.cindoc.csic.es/cybermetrics/articles/v10i1p6.html> accessed 20 Jan 2011.

Shu, W. and Chuang, Y. -H. (2011) 'The Perceived Benefits of Six-Degree-Separation Social Networks', *Internet Research*, 21/1: 26–45.

Smith, A. (1999) 'A Tale of Two Web Spaces: Comparing Sites Using Web Impact Factors', *Journal of Documentation*, 55/5: 577–92.

——. (2001) 'Applying Evaluation Criteria to New Zealand Government Websites', *International Journal of Information Management*, 21/2: 137–49.

Smith, A. G. (2008) 'Benchmarking Google Scholar with the New Zealand PBRF Research Assessment Exercise', *Scientometrics*, 74/2: 309–16.

Spink, A. (2002) 'Introduction to the Special Issue on Web Research', *Journal of the American Society for Information Science and Technology*, 53/2: 65–6.

Thelwall, M. (2001) 'Extracting Macroscopic Information from Web Links', *Journal of the American Society for Information Science and Technology*, 52/13: 1157–68.

——. (2002a) 'A Research and Institutional Size Based Model for National University Web Site Interlinking', *Journal of Documentation*, 58/6: 683–94.

——. (2002b) 'Evidence for the Existence of Geographic Trends in University Web Site Interlinking', *Journal of Documentation*, 58/5: 563–74.

——. (2004) *Link Analysis: An Information Science Approach*. San Diego, CA: Academic Press.

Thelwall, M. and Aguillo, I. F. (2003) 'La Salud De Las Web Universitarias Españolas', *Revista Española De Documentación Científica*, 26/3: 291–305.

Thelwall, M. and Harries, G. (2003) 'The Connection Between the Research of a University and Counts of Links to its Web Pages: An Investigation Based Upon a Classification of the Relationships of Pages to the Research of the Host University', *Journal of the American Society for Information Science and Technology*, 54/7: 594–602.

——. (2004) 'Do the Web Sites of Higher Rated Scholars have Significantly More Online Impact?', *Journal of the American Society for Information Science and Technology*, 55/2: 149–59.

Thelwall, M., Tang, R. and Price, L. (2003) 'Linguistic Patterns of Academic Web use in Western Europe', *Scientometrics*, 56/3: 417–32.

Thelwall, M. and Zuccala, A. (2008) 'A University-Centered European Union Link Analysis', *Scientometrics*, 75/3: 407–20.

Toral, S. L., Martínez-Torres, M. R. and Barrero, F. (2009a) 'Virtual Communities as a Resource for the Development of OSS Projects: The Case of Linux Ports to Embedded Processors', *Behavior and Information Technology*, 28/5: 405–19.

——. (2009b) 'Modelling Mailing List Behaviour in Open Source Projects: The Case of ARM Embedded Linux', *Journal of Universal Computer Science*, 15/3: 648–64.

Toral, S. L., Martínez Torres, M. R. and Barrero, F. (2010a) 'Analysis of Virtual Communities supporting OSS Projects using Social Network Analysis', *Information and Software Technology*, 52/3: 296–303.

Toral, S. L. *et al.* (2010b) 'Current Paradigms in Intelligent Transportation Systems', *IET Intelligent Transport Systems*, 4/3: 201–11.

Vaughan, L. (2006) 'Visualizing Linguistic and Cultural Differences Using Web Co-Link Data', *Journal of the American Society for Information Science and Technology*, 57/9: 1178–93.

Vaughan, L. and Thelwall, M. (2004) 'Search Engine Coverage Bias: Evidence and Possible Causes', *Information Processing and Management*, 40/4: 693–707.

——. (2005) 'A Modeling Approach to Uncover Hyperlink Patterns: The Case of Canadian Universities', *Information Processing and Management*, 41/2: 347–59.

Vigo, M. and Brajnik, G. (2011) 'Automatic web accessibility metrics: Where we are and where we can go', *Interacting with Computers*, 23/2: 137–55.

Vilnai-Yavetz, I. and Tifferet, S. (2012) 'Promoting Service Brands via the Internet: A Survey of the Web Pages of Top Universities', *The Service Industries Journal*, in press, doi: 10.1080/02642069.2011.636423.

Wouters, P., Reddy, C. and Agullo, I. (2006) 'On the Visibility of Information on the Web: An Exploratory Experimental Approach', *Research Evaluation*, 15/2: 107–15.

Yang, B. and Qin, J. (2008) 'Data Collection System for Link Analysis', *Third International Conference on Digital Information Management*, ICDIM, 247–52.

Zhang, J. and Dimitroff, A. (2005) 'The Impact of Metadata Implementation on Webpage Visibility in Search Engine Results (Part II)', *Information Processing & Management*, 41/3: 691–715.