Size of web domains and interlinking behavior of higher education institutions in Europe

Benedetto Lepori*, Isidro F. Aguillo** and Marco Seeber*

**Centre for Organisational Research (CORe), Universitá della Svizzera Italiana, Via Lambertenghi, 10, Lugano, 6900 (Switzerland), benedetto.lepori@usi.ch; phone +41 58 666 46 14; fax +41 58 666 46 19

**Cybermetrics Lab, IPP - CSIC, Albasanz, 26-28, Madrid, 28037 (Spain)

Forthcoming in Scientometrics

Abstract

The aim of this paper is to empirically test whether interlinking patterns between Higher Education Institutions (HEIs) conform to a document model, where links are motivated by webpage content, or a social relationship model, where they are markers of underlying social relationships between HEIs. To this aim, we analyzed a sample of approximately 400 European HEIs, using the number of pages on their web domains and the total number of links sent and received; in addition we test whether these two characteristics are associated with organizational size, reputation, and the volume of teaching and research activities. Our main findings are as follows: first, the number of webpages of HEI websites is strongly associated with their size, and to a lesser extent, with the volume of their educational activities, research orientation, and reputation; differences between European countries are rather limited, supporting the insight that the academic Web has reached a mature stage. Second, the distribution of connectivity (as measured by the total degree of HEI's) follows a lognormal distribution typical of social networks between organizations, while counts of weblinks can be predicted with good precision from organizational characteristics. HEIs with larger websites tend to send and receive more links, but the effect is rather limited and does not fundamentally modify the resulting network structure.

We conclude that aggregated counts of weblinks between pairs of HEIs are not significantly affected by the web policies of HEIs and thus can be considered as reasonably robust measures. Furthermore, interlinking should be considered as proxies of social relationships between HEIs rather than as reputational measures of the content published on their websites.

Keywords. Social relationships, document network, weblinks, higher education.

MSC codes: 62J12, 62P20, 91B74.

JEL classification codes: D85.

1 Introduction

While interlinking patterns between Higher Education Institution (HEI) websites have increasingly been used to characterize their relationships (Thelwall 2012a), their interpretation depends on the understanding of the World Wide Web, as well as of the basic mechanisms that generate links between websites (Gonzalez-Bailon 2009).

Extant literature points to two alternative models. The first model considers the Web as a network of documents, where links are related to the quality of webpage content, like citations in the case of journal papers. From this principle, a growth model of the Web can be derived, in which the connectivity of

websites displays a scale-free power-law distribution, implying that a few websites concentrate most of the connectivity (Barabási, Albert and Jeong 2000; Adamic and Huberman 2000).

An alternative model considers weblinks as the expression of social relationships between respective organizations, associated with factors identified by social network analysis, like social and geographical proximity, access to resources and reputation (Rivera, Soderstrom and Uzzi 2010). In this model, connectivity is expected to follow a lognormal distribution, which is characteristic of the distribution of organizational size in most fields (Ijiri and Simon 1964).

These two models lead to different predictions of distributional properties of weblinks, respectively concerning the factors explaining the observed counts. They also differ in the interpretation of empirical findings – and hence on the potential of link analysis for studies of scientific collaborations. In a document approach, high counts would measure the popularity of web-contents and thus would reasonably be considered as similar to citations in scholarly literature. The use of indicators of web visibility to produce measures of reputation like Web Impact Factors (Almind and Ingwersen 1997; Thelwall 2012b) and Webometrics rankings (Aguillo, Ortega and Fernández 2008) builds on this model.

On the contrary, in a social network approach, the numbers of hyperlinks would measure the strength of the relationships between HEIs. In this perspective, weblinks become a tool to analyze relational structures in higher education, as associated with differences in the level of resources and with reputational hierarchies between HEIs (Lepori, Barberio, Seeber and Aguillo 2013).

The expected impact of HEI web policies on the connectivity of their websites is also different; according to the document model, HEI publishing on the web is more useful and high-quality content would receive higher counts of weblinks, whereas in the social relationship model published content is expected to have a limited impact on connectivity, once a baseline level of Web presence is achieved.

The aim of this paper is to contribute to the clarification of this issue through the statistical analysis of interlinking patterns in a large sample of European HEIs.

In the first step, we analyze the web policies of HEIs through a comparison of the size of their web domains (measured through the number of webpages) with organizational characteristics (size, research orientation, reputation) to test whether systematic differences across HEIs and countries found in previous studies, are still observed (Payne and Thelwall 2007).

In the second step, we analyze the distributional properties of weblinks between HEIs and test whether the total number of links sent and received by an individual node (*degree*), as well as the count of links between two HEIs, can be predicted from organizational characteristics, the number of pages on the website, as well as their geographical and social proximity.

2 Framework and research questions

Owing to the increasing use of the World Wide Web as a communication media tool in academia, the analysis of the distribution of links between HEIs has become increasingly popular (Bar-Ilan 2009; Thelwall 2012a).

In this respect, weblinks have potential advantages against other measures of relationships, such as co-authorships (Glänzel and Schubert 2005) and collaborations in international research projects (Heller-Schuh, Barber, Henriques, et al 2011). First, they provide broader measures covering other academic activities and thus can be extended beyond the core of research-intensive universities; second, counts of weblinks are sufficiently high to provide data for smaller HEIs; finally, despite technical issues with data

collection related to the structure of websites (Thelwall and Sud 2011), weblink data is relatively easy to retrieve, even in very large samples.

A first set of studies descriptively analyzed the structure of the network between HEI websites – highlighting patterns such as the importance of national networks (Ortega, Aguillo, Cothey and Scharnhorst 2008) and the emergence of a global network between the most reputed HEIs (Lee and Park 2012). Interlinking patterns between HEIs have also been shown to differ substantially from networks derived from co- authorships (Kretschmer, Kretschmer and Kretschmer 2007).

A second set of studies tested whether the number of links between two HEIs can be predicted from organizational characteristics like size and reputation (Thelwall 2002a; Vaughan and Thelwall 2005), as well as geographical distance (Thelwall 2002b); these studies provide statistical support that the presence and number of links between HEIs displays regular patterns associated to their size, reputation, and geographical proximity (Seeber, Lepori, Lomi, Aguillo and Barberio 2012).

Complementarily, micro-level studies investigated the motivations and mechanisms behind hyperlinking, showing that most links between HEIs are associated with academic activities, but only a relatively small percentage are directly research-associated (Bar-Ilan 2005; Vaughan, Kipp and Gao 2007). These studies support the assumption that weblinks might provide useful information on HEI relationships, but with a broader scope in terms of the activities covered than publication-related measures (Wilkinson, Harries, Thelwall and Price 2003).

2.1 Document vs. organizational models of the web

The interpretation of links depends on the underlying model, specifically concerning how they are generated on the Web and which relationships they bare to the organization who owns the website.

Two models can be compared, namely a document network, and a social relationship model (Gonzalez-Bailon 2009). They foresee different generative mechanisms for weblinks, and accordingly lead to diverging predictions concerning their distributional properties, as well as the association with organizational characteristics.

In a document model, its oldest and most popular interpretation, the Web is conceived as a set of documents connected by hyperlinks (Barabási, Albert and Jeong 2000); the motivations for interlinking are related to the relevance of webpage content in the linked page, and only indirectly to the characteristics of their authors. Very much like bibliographical citations, weblinks can be interpreted as voting procedures concerning the quality and usefulness of web content attributed by other users (Thelwall 2006; Gonzalez-Bailon 2009).

It has been shown that if a *preferential attachment* principle applies, where the probability that a newly created document links to an existing one, is based on extant level of centrality, the growth of the Web would lead to the power-law distribution of the links received, where only a few documents account for most of the connectivity in the network (Barabási, Albert and Jeong 2000; Katz and Cothey 2006). Highly popular content-based websites like Wikipedia have been shown to conform to this model (Capocci, Servedio, Colaiori, et al 2006). Outside the web, the preferential attachment model applies to networks of scientific co-authorships, following the cumulative nature of scientific reputation (Wagner and Leydesdorff 2005; De Stefano, Fuccella, Vitale and Zaccarin 2013).

Empirical studies provide support to this model for the web as a whole (Adamic and Huberman 2000), which is indeed characterized by a few very popular websites with a disproportionate number of links. However, other studies suggest that the distribution of connectivity in more restricted domains, like

industrial sectors, universities, and newspapers, display a less skewed distribution (Pennock, Flake, Lawrence, Glover and Giles 2002).

These findings suggest that, at least in domains where the Web is characterized mostly by organizational websites whose main aim is to provide visibility to the organization's core activities, a social network model might provide a more adequate representation of interlinking patterns.

In this model, a link between two websites is a marker of an underlying social relationship between the respective actors. In this model, the content of the related webpages plays a less relevant role for establishing weblinks, consistently with the fact that a large share of weblinks from university websites originate from either institutional pages or from compilations of links (Bar-Ilan 2005).

The social network literature identifies a number of underlying motivations in establishing a social relationship, including connecting preferentially to actors which hold important material resources and status (Cattani, Ferriani, Negro and Perretti 2008), geographical and social proximity (belonging to the same social group, region or language; Thelwall, Tang and Price 2003), and homophily, i.e. linking preferentially with actors who display similar characteristics (for example HEIs specialized in the same scientific domain; Rivera, Soderstrom and Uzzi 2010).

In most social networks, preferential attachment – i.e. linking preferentially to well-connected people - is an important mechanism to establish ties, as it is usually considered that central people hold more resources and status (Owen-Smith and Powell 2008). However, unlike in document networks, material resources have a direct impact on establishing ties (Gonzalez-Bailon 2009) which limits the level of asymmetry in the network. Since the volume of activities is expected to be roughly proportional to size, the total number of links sent and received should display a distribution similar to the one of organizational size, which in most sectors is known to follow a lognormal distribution (Gibrat's law; Ijiri and Simon 1964). Moreover, unlike document networks, social networks are subject to social norms of reciprocation, i.e. a general principle that if one actor links to another, the latter is expected to reciprocate the relationship, which limits the extent to which social networks can become asymmetrical despite differences in status. Keeping some level of reciprocation is known to be a general feature of many social networks, and more generally of the whole society (Gaudeul and Giannetti 2013).

Thus, while social networks indeed display large differences in the centrality of individual actors, they are expected to show a lower concentration of connectivity and a higher degree of reciprocation than document networks, purely based on the preferential attachment principle.

2.2 Web policies and interlinking patterns

Models of the web can also be associated to different approaches concerning the Web policies of organizations. In a document model, the WWW might become a central competitive asset to attract potential audiences and to market their product, a phenomenon well known in domains like publishing and electronic commerce. Some organizations are then expected to invest heavily in web publishing, as related to specific content, leading to higher visibility when compared to what would be expected from their size and resources.

On the contrary, in a social relationships model, the visibility an organization achieves on the Web is essentially associated with its size, reputation, and resources; meaning the main focus of the WWW would be to provide a fair presentation of activities, while a convergence towards a standard level of web presence is expected.

There is some evidence that HEI websites present features of both models, but with a prevalence towards the organizational one. In many HEIs, the design and structure of websites are centralized by the informatics or communication department, who also provide guidelines and templates for content produced by departments and individuals (Peterson 2013). At the same time, the Web site structure mirrors the organizational model, with most of the content located on the various departmental websites or personal webpages. The combination of central templates – which are increasingly supported by the introduction of Content Management Systems – and the decentralized production of content, leads to the expectation that the number of webpages is related to the number of academic staff. For instance, personal webpages account for an increasing share of HEI websites, but in many HEIs there are standards on how pages care organized and the number of subpages allowed (even if published content might display strong variations concerning quality and detail; Más-Bleda and Aguillo 2013).

On the other hand, some HEIs increasingly use the Web for marketing purposes and for attracting students (Astani 2013), as well as for providing on-line educational resources; libraries and open access repositories of academic content are potentially other areas where HEIs might invest in order to increase their visibility by providing additional content, even if the available data shows that repositories account for a substantial share of webpages in only a limited number of HEIs (Aguillo, Ortega, Fernández and Utrilla 2010). Accordingly, some level of variation in the size of Websites might be related to these phenomena.

Different models of interlinking generation also hold different predictions concerning the impact of HEIs' web policies.

On the one hand, in a document model, given the power-law distribution of connectivity, few webpages in a website are expected to account for most of the links received, as related to the specific characteristics of their content. Accordingly, we expect that the total number of links sent and received is only loosely associated with organizational characteristics, but also with the number of pages published on the web.

On the contrary, a social network model foresees that the level of connectivity of an HEI is basically related to its resources, reputation, and position in the academic world, while characteristics of the website would have limited impact (at least concerning the academic web, to which our study is limited).

In this model, the size of the web domain is expected to have a moderate impact on the links sent and received (once we control for size, education, research, etc.). Namely, counts of weblinks between two HEIs are the aggregate outcome of a micro-process of interlinking between two individual webpages, whose probability is associated with organizational characteristics (Seeber, Lepori, Lomi, Aguillo and Barberio 2012). However, when a link is established, a larger number of webpages might increase the counts (as there might be multiple links related to the same underlying relationships). Hence, the size of the web domain is expected to increase the counts of weblinks between HEIs, which according to their characteristics and proximity are already likely to be connected.

The provision of additional educational content, as well as repositories and academic content on personal webpages, is not expected to significantly affect these patterns. On the one hand, educational content is mostly directed to students and external audiences and thus is not likely to strongly affect academic connectivity. On the other hand, expectations of additional interlinking due to citations among scientists have not taken into account the success of permanent URLs (pURLs) like DOIs, a unique identifier that is independent of academic web domains. This lack of proper recognition also affects institutional repositories where the use of handles is frequently recommended; another type of pURLs that can or cannot include the target web domain of the HEI.

2.3 Research questions and hypotheses

We empirically investigate this issue in three different steps.

First, we analyze to which extent the size of web domains (measured by their total number of webpages) is associated to HEI size, as well as to their volume of activities and academic reputation. While some level of association is expected, we expect it to be much stronger if the WWW is a web of social relationships rather than of document relationships.

Second, we investigate the distributional properties of the total number of links sent (*outdegree*) and received (*indegree*) by the HEIs in our sample. In a document model, indegree is expected to display a power-law distribution, where few HEIs receive most of the links. Moreover, the network is expected to be highly asymmetric, with most HEIs sending more links than they receive and a few very central sites receiving most of the links. On the contrary, in a social network model, the network should be more symmetric, as reciprocity is known to be a central feature of most social networks. Moreover, both indegree and outdegree are expected to be closely associated with organizational characteristics and reputation, and share their distributional properties.

Third, we test the predictive ability of different HEI characteristics, including the size of their web domains, on the total degree and on the number of links between two HEIs. In a social network model, we expect the existence and strength of the relationship to be largely predicted by these characteristics, while the size of the web domain should have an impact on the counts of links, but not on the existence of a connection.

3 Methodology

Three data sources were used in this study: The EUMIDA project (2009) provided organizational data, including the size of the institution, reputation, and geographical position (Lepori and Bonaccorsi 2013); bibliometric data to measure international reputation was derived from Scimago Institutions ranking (http://www.scimagoir.com/), whereas the Cybermetrics Lab collected the web size and interlinking data by processing information from MajesticSEO (http://www.majesticseo.com/, February2013).

3.1 Sample

Our sample includes a set of 396 European universities from 20 countries[1] constructed as follows: all HEIs in the EUMIDA database, which are also covered by the Scimago Institutions Ranking for the year 2011, were included. These HEIs had at least 100 Scopus publications in the year 2009. The countries with the largest number of cases were the UK (86), Germany (66), Italy (58) and Spain (47); several Polish universities had to be excluded because their shared web domains made the web data obtained unreliable.

It is important to analyze coverage in respect to the full EUMIDA database, which provides a nearly complete view of higher education in Europe (EU-27 plus Norway and Switzerland, less France). Our sample includes only 16% of the 2,457 HEIs in the database, and 29% of the 1,378 HEIs labeled as research-active (accounting for the largest share of activities, students and research). The sample HEIs include 59% of students, 73% of doctoral students, and 70% of staff in research-active EUMIDA HEIs, revealing that these HEIs are larger and more research-oriented.

As a matter of fact, our sample includes half (392 out of 847) of doctorate-awarding HEIs in the EUMIDA database; the sample is then quite representative of doctorate-awarding universities, but not of other

[1] Following countries were covered: Austria, Bulgaria, Switzerland, Germany, Estonia, Spain, Finland, Hungary, Ireland, Italy, Lithuania, Latvia, Netherlands, Norway, Poland, Romania, Sweden, Slovenia, Slovakia and UK.

types of HEIs. Additionally, they account for 87% of the total degree (total links sent and received) in the 1181 HEI sample used in Seeber, Lepori, Lomi, Aguillo and Barberio 2012; in other words, HEIs in our sample account for most of the connectivity in European academic websites.

3.2 Interlinking data and data on the size of web domains

Traditionally, web data for very large populations has been extracted from general search engines such as Google, Yahoo, or Bing. The size and global coverage of their databases made them suitable for webometric studies, but unfortunately their capabilities and operators for link analysis were limited. Additional problems related to the opaque and irregular behavior make it advisable to search for alternative sources.

Since 2010, the Cybermetrics Lab has evaluated three different link data commercial providers: the US SeoMoz (http://www.opensiteexplorer.org/) the British MajesticSEO, (http://www.majesticseo.com/) and the Ukrainian ahrefs (http://ahrefs.com/) services.

The API provided by MajesticSEO was especially suited for obtaining both web size (number of different URLs indexed) and linking information among pages or complete web domains. Two main problems persisted in the data collection using MajesticSEO: The first one is related to duplicate domains for 26 universities. The second problem is that the system provides erroneous results for universities with shared domains, a common situation for several Polish universities that use city-related domains (for example Wrocław University of Technology / Politechnika Wrocławska both use the same city domain: pwr.wroc.pl). In the first case, the procedure was to combine (addition) the link numbers to both domains, but as it can be assumed a large overlap exists between the contents of both domains, the web size was obtained from the maximum value. In the second case, all Polish universities in those circumstances were excluded from the analysis. However, the universities included from this country are the most prestigious and productive.

From this source, we derive the following variables for our analysis: the *number of weblinks* from one HEI in the sample to the other, the *total degree* for each HEI in the sample (sum of all links sent and received from other HEIs in the sample), and finally the *size of the web domain* from each HEI.

3.3 Organizational data

For the HEIs in the sample, the following data has been used. *Size* of the HEI, as measured through the number of academic staff, *educational activities* as the number of undergraduate students (students at level 5 of the International Standard Classification of Educational Degrees; ISCED), *research intensity* as the ratio between the number of PhD students (students at ISCED 6 level) and undergraduate students. We use as a measure of *international reputation* the brute force indicator calculated as the product of the total number of publications of an HEI with their average impact factor normalized by the number of academic staff (van Raan 2007); this indicator builds on the insight that the international visibility of an HEI is related both to quality and the volume of output.

When predicting counts of weblinks between two HEIs, we include a number of variables that characterize their similarity and proximity in physical and social spaces (Rivera, Soderstrom and Uzzi 2010). These include a measure of *subject similarity* between two HEIs, based on an overlap in the distribution of students from educational fields. Furthermore, a measure of *kilometric distance* between two HEIs, which has been constructed from the geographic coordinates of the HEI's web domain IP, and finally a dummy variable if two HEIs belong to the same *country*.

3.4 Statistical methods

Our empirical strategy is twofold. On one hand, we focus on the level of individual HEIs, determining to which extent the size of their web domains and the total number of links sent and received from other HEIs in the sample (*total degree*), can be predicted by using their organizational characteristics. Thus, the number of observations in this analysis is 396.

On the other hand, we analyze the number of weblinks from one HEI to another in our sample, testing whether they are predicted by characteristics of the sender and receiver HEIs (including the size of the web domain), as well as by their similarity and proximity. The number of observations in this analysis is 396 * 395, i.e. the number of directed pairs in our sample.

The two analyses share a number of independent variables, while their dependent variables are closely related, as total degree is the sum of all links sent and received with other HEIs in the sample. However, different distributional properties lead to the choice of different statistical methods.

a) *Size of web domain and total degree*. We first provide descriptive statistics concerning HEI size, web domain size, and total degree, in order to analyze their distributional characteristics, and to identify significant correlations, as well as outliers.

Since descriptive statistics display the distribution of size, and the web domain size and degree in our sample is nearly lognormal, we test relationships between them using OLS regressions with a log-log transform of all variables.

b) *Counts of weblinks between pairs of HEIs*. After providing descriptive statistics on the distribution of weblinks, we perform regressions between link-counts (dependent variable) on the one side, and organizational characteristics of HEIs and size of web domains on the other.

Since we work with count data, we use a negative binomial regression which includes a parameter to model over-dispersion as compared to a standard Poisson regression (Cameron and Trivedi 1998). This type of model is robust against the non-normality of distributions and the presence of outliers, and has successfully been used to model web links (Seeber et al. 2012). Given the large number of 0's in our sample, we employ a hurdle model, which specifies a separate model (a logistic regression) for predicting zeros. If the threshold of the hurdle is reached, the case is passed to the negative binomial model, which predicts the expected count of weblinks.

4 Results

4.1 Descriptive statistics

Descriptive statistics show that the size and total degree of the web domain are strongly skewed and display several very high values (Table 1). Whereas on average, HEIs in the sample have 180,000 webpages and nearly 10,000 weblinks sent and received, there are four HEIs with more than 1 million webpages (Slesian University of Technology, University of Vigo, University of Bern, University of Amsterdam) and two HEIs with a degree above 100,000 links (Aachen and Regensburg).

	Mean	Median	Min	1Q	3Q	Max	STDEV
Size	1,518	1,256	63	797	2,020	6,571	1,017
Size web domain	179,683	124,151	3,871	70,904	218,603	2,414,850	213,514
Total degree	10,212	5,894	37	2,512	12,475	269,652	17,342
Research intensity	.06	.04	.00	.02	.08	.47	.06
Reputation	4.88	4.00	.17	1.86	6.35	41.77	4.59

Table 1. Descriptive statistics of organizational variables, web domain size and total degree

The distribution of HEI size is lognormal, as expected from Gibrat's law of organizational size (Ijiri and Simon 1964; Kolmogorov-Smirnov statistics: .470, p=.980); the same applies for the size of the web domain (Kolmogorov-Smirnov statistics: .796, p=.551). Total degree is slightly more skewed (Kolmogorov-Smirnov statistics: 1.412, p=.037), but nevertheless much nearer to a lognormal distribution than to a power-law distribution (Figure 1).

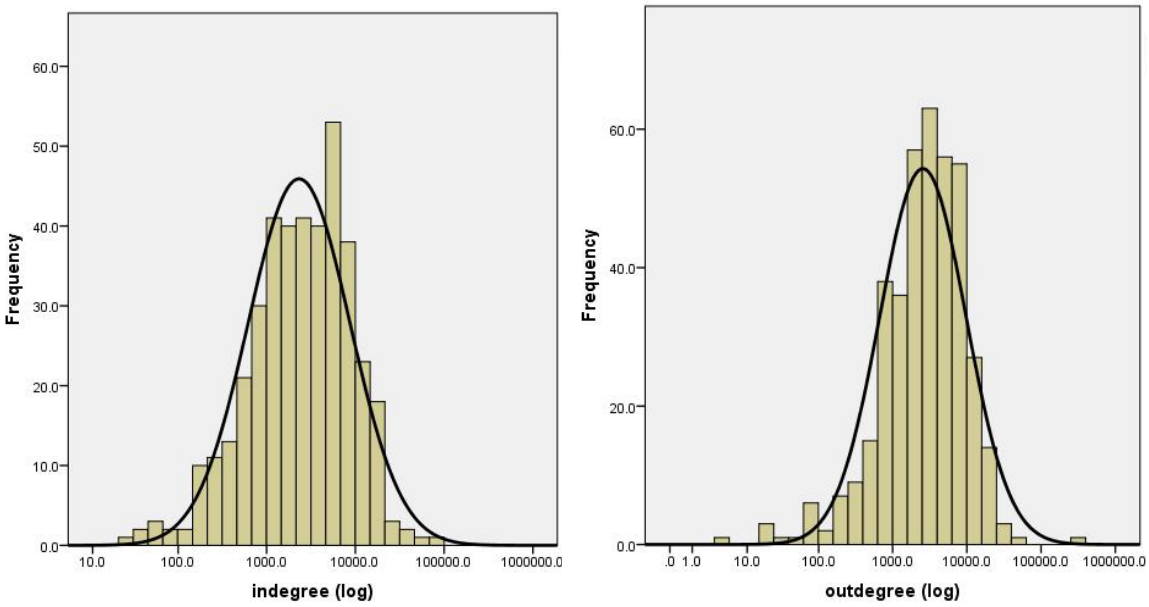

Figure 1. Distribution of indegree and outdegree

A more detailed analysis of degree does not reveal the highly asymmetric winner-take-all structure expected for document networks. Namely, the correlation between indegree and outdegree is very high (.858** on a log-log scale), whereas indegree is actually more skewed than outdegree (skewness 6.6 against 15.4). While there is some concentration of links, this is not extremely high as would be predicted by a power—law distribution: the Gini coefficient is only moderately high (0.57 for indegree and 0.58 for outdegree), whereas the 5% of HEIs with the highest number of links account for only 28% of the indegree and 32% of the outdegree. Also, ln_size is correlated .628** to ln_websize and .630** to ln_degree, whereas ln_degree and ln_websize are correlated to .818**.

These results provide preliminary evidence that the distribution of degree displays distributional properties that are fairly similar to organizational size, but which depart substantially from a power-law distribution.

4.2 Web domain size and degree

Given their relationship with size, we analyze the distribution of web domain size and degree using normalized variables against HEI size, since this will allow for the identification of HEIs with much larger websizes (respectively number of weblinks) than expected.

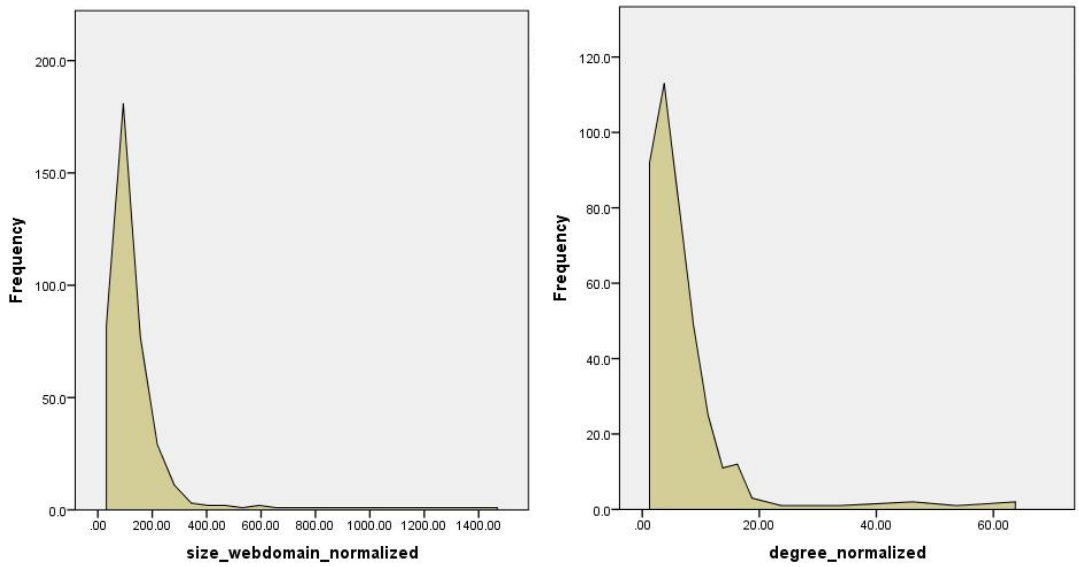

Figure 2. Distribution of the size of web domain and total degree normalized

N= 396

As shown by Figure 2, the distribution of size of the web domain and the total degree once they are normalized by size is rather regular, except for a long queue of HEIs displaying much larger values than the average.

Differences between countries concerning both the number of webpages and the degree are smaller to those found in previous studies (Thelwall, Binns, Harries, Page-Kennedy, Price and Wilkinson 2002, Payne and Thelwall 2007).

Convergence is stronger concerning the number of webpages: as a matter of fact, only 5% of the variance in the size of the web domain (normalized) is between countries (anova one-way test). Differences in country medians are also rather limited for most countries, with 15 out of 20 countries having a median between 80 and 140 pages per units of academic staff, and only two Eastern European countries displaying much lower numbers (Bulgaria 58 and Romania 46 pages per unit of academic staff). These findings support the insight that the diffusion of web communication is a general trend which has spread across all of Europe.

To some extent, organizational and technical changes may have fostered this tendency towards uniformity: on the one hand, for many HEIs, their websites are centrally managed and the design and content is created, modified, and uploaded by a small group following strict rules. On the other hand, the generalized use of Content Management Systems (CMS like Drupsal, Wordpress, joomla) makes academic websites more homogeneous with very similar patterns. Probably these tools have limited the number of new links by the occasional author, which is now a large group of current academic web editors. Of course, this does not imply that cultural differences concerning the look and feel of Websites is disappearing (Callahan 2005), but this pertains more to how the information is presented than to the organization of websites.

In the European context, European policies towards the European research area are likely to have contributed to this convergence, by emphasizing the importance of communication, but especially by fostering the imitation of practices from the leading universities in Europe. Even at the global level, some commonalities emerge between university websites, such as the widespread use of English as their first or second language (Callahan and Herring 2012).

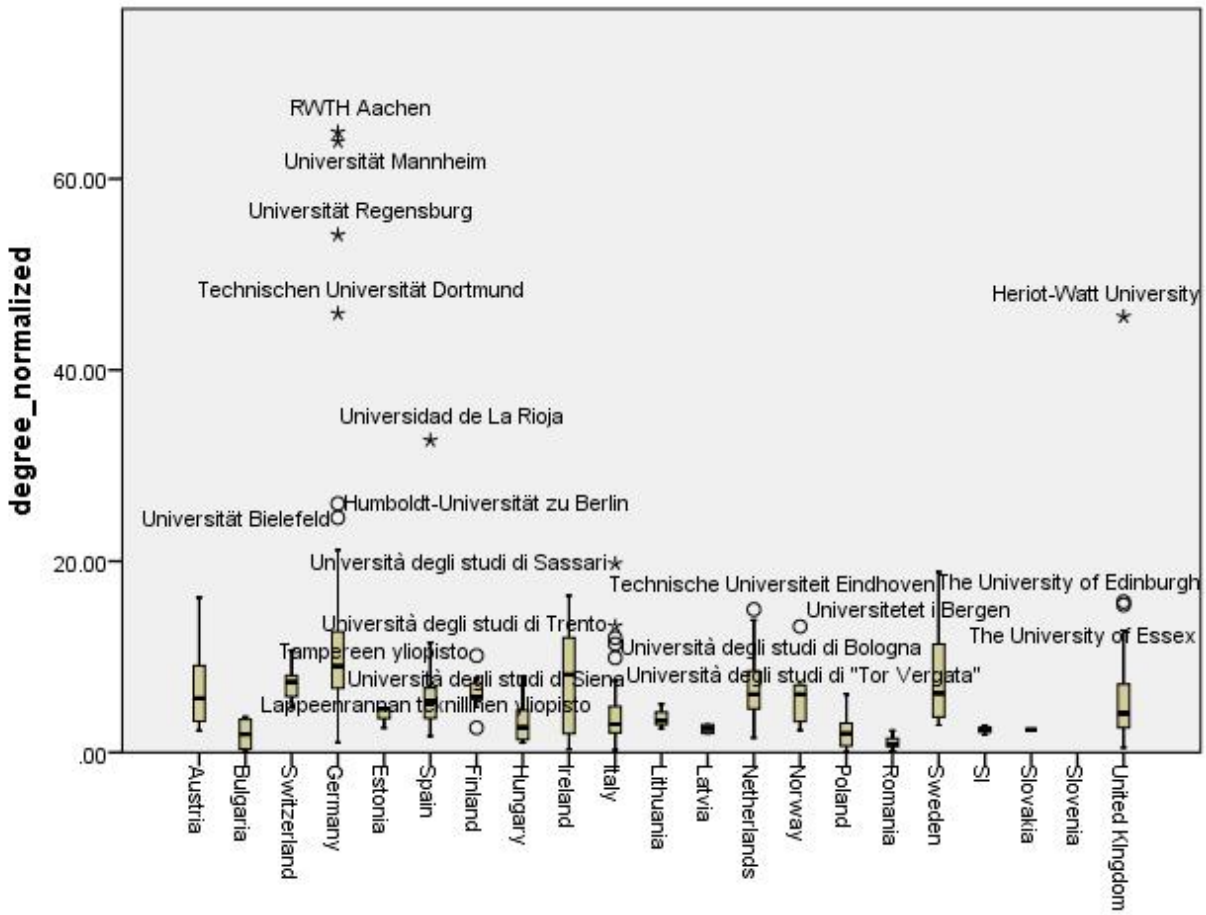

Figure 3. Distribution of degree normalized by country

As shown by Figure 3, the situation is slightly different concerning total degree, where 24% of the variance is between countries. As a matter of fact, Eastern European countries display systematically lower levels of interlinking than Western European countries, accounting for their lower level of centrality in the European University network (Ortega, Aguillo, Cothey and Scharnhorst 2008). Differences in this respect are much larger than those concerning the size of web domains (the median of count weblinks for Bulgarian HEIs is about 1/5 than for German HEIs).

The analysis of outliers is informative of the mechanisms which might lead HEIs to have a very large number of webpages (in respect to their size), respective of incoming and outgoing weblinks. We identify 25 outliers for the size of the web domain and 20 for total degree.

The 25 outliers for the size of the web domain do not display particular patterns concerning country distribution, size, research intensity, and international reputation. Among them, we find both small and specialized HEIs (Silesian University of Technology, London Business School), as well as large generalist universities, like Amsterdam and Bern. At least in one case (University of Rioja), the high number of webpages is due to a bibliographic database, which accounts for 97% of the webpage counts. Importantly, only 4 of these HEIs are outliers concerning their total number of links as well, thus publishing a number of webpages much higher than expected from size does not translate into a much higher level of connectivity. While a more in-depth investigation of these cases would be required, we foresee a few mechanisms which could explain these outliers: the presence of large repositories generating a high number of webpages,

dedicated websites to external audiences (for example providing educational materials) – however with a limited impact on academic connectivity – and finally, indexing mistakes by search engines.

The outliers in terms of total degree display different patterns: they are all in Western European countries (12 out of 20 in DE), they display a higher level of international reputation than average HEIs in the sample, and they tend to have a larger number of webpages per unit of academic staff. It is known from previous work that the most extreme cases (like RWTH Aachen) are explained by the presence of large web repositories. In other cases, outliers seem to be generated by the combination of a strong international reputation, belonging to well-connected countries, and a communication policy more active than the average HEI in the sample. A few cases might also be generated by an extremely high count between individual HEIs, such as the cases of the University of Dortmund and Regensburg (about 40,000 links sent, i.e. 40% of their total degree). While an in-depth analysis of these cases might be interesting, their number is sufficiently low and they can readily be identified from a distributional analysis.

4.3 Regression models for web domain size and degree

In a second step, we test whether the size of the web domain, respectively the total degree, can be predicted by organizational attributes.

Correlations between the log-transformed attributes are rather small, the largest is between ln(size) and ln(students) with .351** and between ln(research intensity) and ln(reputation) with .341**. Our choice of independent variables is then suitable to limit collinearity problems between size, educational, and research activities.

	Size only		All variables			Excluding outliers		
	Estimate	SE	Estimate	SE	Beta	Estimate	SE	Beta
Intercept	5.758***	.370	6.486***	.493		4.903***	.430	
Ln size	.836***	.052	.729***	.055	.550	.805***	.048	.633
Ln research intensity			.084*	.041	.085	.097*	.034	.084
Ln reputation			.226***	.041	.224	.230***	.034	.246
Ln students			.106*	.050	.089	.097*	.042	.087
		Df		Df			Df	
Adjusted Rsquare	.397	395	.453	394		.584	369	
Residual mean sum of squares	.494	394	.448	390		.301	365	
F statistics	259.891***	1	82.764***	4		127.982***	4	

Table 2. Size of web domain: regression results

Dependent variable: Ln(web domain_size).

The model including only size provides a good fit for the size of the web domain; the other attributes (educational activities, research intensity, reputation) are significant as well, consistently with the expectation that a stronger orientation towards research implies the production of additional content and that more reputed HEIs also have stronger incentives to publish Web content since they expect higher levels of visibility. However, their contribution in predicting the size of the web domain is limited (Table 2).

Excluding the 25 outliers in terms of the size of the web domain normalized by size, improves the level of fit, but does not significantly affect the estimate.

We conclude that HEI size is a highly significant predictor of the size of the web domain; since its coefficient is significantly lower than 1, the number of webpages grows at a slower rate than academic staff, but the difference with a linear relationship is not very large (doubling the HEI size would increase the number of webpages by about 80%). We notice that despite this common trend, a substantial share of variation in web domain size remains which is not explained by organizational size and activities, suggesting that some HEIs provide additional content on the web, for example through repositories or educational websites.

	Size only		All organizational characteristics		Web domain size, without size		All variables		
	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Beta
Intercept	.216	.523	-.373	.632	-4.909***	.564	-4.363***	.487	
Ln size	1.172***	.073	.946***	.071			1.174***	.055	.629
Ln research intensity			.153**	.052	.103*	.040	.083*	.039	.059
Ln reputation			.527***	.052	.334***	.042	.336***	.041	.237
Ln students			.211**	.127	.194***	.048	.122**	.049	.073
Ln web domain size					.983***	.042			
Ln web domain size normalized							.843***	.049	.471
		Df				Df		Df	
Adjusted Rsquare	.393	395	.547	394	.726	394	.742	394	
Residual mean sum of squares	.984	394	.736	390	.463	390	.418	389	
F statistics	256.452***	1	119.904***	4	261.605***	4	228.192***	5	

Table 3. Total degree: regression results

Dependent variable: LN(degree).

Concerning degree, size remains quite an important predictor, explaining about 40% of the total variance in the sample (Table 3). Educational activities and research intensity are also significant and their contribution to predicting degree is higher than for webpages, consistently with the hypothesis that interlinking behavior is related to HEI activities and is selective towards research-oriented and internationally reputed HEIs.

In turn, the introduction of the size of the web domain significantly improves the predictive ability of the model: having a larger website translates into an HEI sending and receiving significantly more links, even if the impact of this factor remains smaller than the one of size. Results do not change significantly when excluding outliers for the size of the web domain, as well as for degree. Moreover, separate regressions for indeegree and outdegree provide essentially the same results.

For all models, robustness checks concerning collinearity (variance inflation factors), heteroskedasticity (Durbon-Watson statistics) and normality of residuals provide satisfactory values, thus supporting the robustness of findings.

In order to assess the validity of results on a larger sample of European HEIs, we performed a test using weblink data derived from Seeber, Lepori, Lomi, Aguillo and Barberio 2012, while for the size of the web domain we used data collected in 2011 from Google.

Findings confirm those presented: first, the size of the web domain is predicted quite well from the HEI size (coefficient 1.219*** for ln(size) and Rsquare .579, N=1210); second, total degree can be predicted from HEI size and web domain size with high precision (coefficients 1.102*** for ln(size) and .629*** for ln(web domain size normalized), Rsquare .753; N=1138). Both the coefficients and the level of fit are quite similar to those presented in Table 2 and Table 3.

Besides hinting that the findings presented are likely to be valid for European HEIs in general, beyond the research-oriented universities considered in this paper, these results also support their robustness, since they have been generated from data collected at different points in time (2011 instead of 2013) and from different sources (google instead of Majestic).

4.4 Predicting weblinks

As a third step we use the same variables, plus a number of variables describing similarity and proximity between HEIs, to predict weblink-counts between pairs of HEIs. In order to simplify readability, we measure

size and the size of web domains in 1,000 units throughout this section. For distance, we used a log transform as this is known to better predict the impact of geographic distance on the strength of relationships (Daraganova, Pattison, Koskinen, et al 2012).

Counts of weblinks display the well-known skewed distribution (Table 4). More than half of the dyads have no link, while slightly less than 3,000 dyads (less than 2% of the total) have more than 99 links and account for more than half of total links.

Counts of weblinks	**0**	**1-9**	**10-99**	**100-999**	**1000-**	**Total**
Number of dyads	80,706	50,289	22,512	2,784	129	156,420
Sum of links	0	195,085	671,261	624,284	378,617	1,869,247

Table 4. Distribution of weblinks by categories

This distribution has substantial implications for our analysis, as it implies that total degree for individual HEIs will be determined by a few strong connections, representing a small share of the total dyads.

Since the specific focus of this paper is on investigating the impact of the size of the HEI web domain on connectivity, we compare three models: the baseline model tested in Seeber, Lepori, Lomi, Aguillo and Barberio 2012, a model where we replace organizational size with the size of the web domain, and a model where web domain size is introduced alongside organizational size. We compare the three models concerning overall fit, strength of the effects, and predictions of counts.

	size 1000 model			webdomain model			full model		
	Negative binomial model								
	Estimate	SE		Estimate	SE		Estimate	SE	
(Intercept)	2.124	0.072	***	1.690	0.067	***	1.236	0.0693	***
size_sender1000	0.285	0.006	***				0.329	0.0064	***
size_receiver1000	0.356	0.006	***				0.411	0.0064	***
reputation_sender	0.017	0.002	***	0.008	0.002	***	0.010	0.0019	***
reputation_receiver	0.053	0.002	***	0.039	0.002	***	0.046	0.0019	***
res_int_sender	2.927	0.139	***	3.193	0.159	***	2.741	0.1346	***
res_int_receiver	3.423	0.136	***	3.719	0.140	***	3.282	0.1310	***
webdomain_sender_1000				0.0002	3.645E-06	***			
webdomain_receiver_1000				0.0028	5.127E-05	***			
webdomain_sender_normalized							0.0020	0.0001	***
webdomain_receiver_normalized							0.0021	0.0001	***
subject	0.589	0.028	***	0.752	0.028	***	0.565	0.0264	***
log_distance	-0.889	0.021	***	-0.724	0.022	***	-0.788	0.0196	***
country	1.627	0.020	***	1.645	0.022	***	1.580	0.0194	***
Log(theta)	-1.247	0.014874	***	-1.231	0.028	***	-1.131	0.0138	***
	hurdle model								
	Estimate	SE		Estimate	SE		Estimate	SE	
(Intercept)	-0.640	0.085	***	-0.020	0.083		-1.151	0.087	***
size_sender1000	0.623	0.007	***				0.650	0.007	***
size_receiver1000	0.546	0.007	***				0.571	0.007	***
reputation_sender	0.010	0.001	***	0.007	0.001	***	0.010	0.001	***
reputation_receiver	0.047	0.001	***	0.045	0.001	***	0.047	0.001	***
res_int_sender	3.185	0.123	***	4.323	0.119	***	3.034	0.124	***
res_int_receiver	5.320	0.129	***	6.284	0.125	***	5.182	0.129	***
webdomain_sender_1000				0.0002	5.521E-06	***			***
webdomain_receiver_1000				0.0017	3.499E-05	***			***
webdomain_sender_normalized							0.0010	4.301E-05	***
webdomain_receiver_normalized							0.0009	4.273E-05	***
subject	2.268	0.026	***	2.503	0.026	***	2.282	0.027	***
log_distance	-1.066	0.026	***	-0.986	0.025	***	-1.006	0.026	***
country	1.645	0.029	***	1.574	0.028	***	1.705	0.029	***
Theta:	0.2873			0.292			0.3227		
Number of iterations in BFGS optimizatic	24			25			29		
Log-likelihood:	-3.48E+05	on 21 df		-3.52E+05	on 21 df		-3.47E+05	on 25 df	

Table 5. Count of weblinks. Results of the hurdle negative binomial model

Top part of the table: negative binomial model for predicting non-zero counts. Bottom part of the table: hurdle model for predicting zeros (logistic regression model).

a) *Model fit*. As shown by Table 5, the baseline model reproduces the results obtained in Seeber, Lepori, Lomi, Aguillo and Barberio 2012, thus confirming that our sample is sufficiently representative of HEI connectivity in the European academic web. Strikingly, the model where HEI size has been replaced by the size of the web domain provides a lower level of fit. In other words, the number of webpages is a less accurate predictor of the number of weblinks than organizational size (number of academic staff), a result which supports the superiority of a social network model.

Expectedly, the model including both measures provides the best fit. The difference is statistically significant (log-likelihood ratio test, $p<0.001$), thus showing that nevertheless the web domain's size provides relevant information for connectivity; however the coefficients of the other variables are only slightly modified.

b) *Strength of the effects*. Since the coefficients of negative binomial regressions are multiplicative, the strength of the effects of independent variables is expressed as a percentage change of the baseline value. While confirming previous findings that belonging to the same country has by far the strongest impact on connectivity, followed by organizational size, results show that the impact the number of webpages on the likelihood of two HEIs being connected is relatively limited, while it is significantly larger on the expected counts of links between two HEIs (once they are connected by at least one link; Table 6). This supports our interpretation that increasing the number of webpages tends to multiply existing links, but not to generate new connections.

Variable	**Range**	**Expected percentage change**	
		Likelihood of linking (odd ratios)	Expected count (predicted means)
Sender size (1000)	1.01	93%	39%
Receiver size (1000)	1.01	78%	51%
Reputation sender	4.59	5%	5%
Reputation receiver	4.59	24%	23%
Research intensity sender	0.056	19%	17%
Research intensity receiver	0.056	34%	20%
Websize sender normalized	143	16%	33%
Websize receiver normalized	143	14%	35%
Subject	0.25	77%	15%
Log distance	0.37	-31%	-25%
Country	1	450%	385%

Table 6. Impact of independent variables on number of weblinks

Range for continuous variables is their standard deviation.

c) *Prediction of counts*. Results confirm the baseline model's sufficient ability to classify dyads by classes of strengths: overall the model is able to correctly identify 78% of the zero dyads and 68% of the non-zero ones. Predictive ability remains quite good for identifying dyads above 9 links (70%), and even 99 links (41%); the predictive ability is much lower for cases above 999 links (9%), since these are very rare and might be generated by specific conditions which cannot be captured by organizational characteristics.

The model including only the size of the web domain is significantly inferior to the one with organizational size (60% against 56% of correctly classified cases in four classes of intensity); the predictive ability is strikingly inferior in identifying cases above 9 links (19%), respectively 99 links (2%). In other words, the size of the web domain is a much less reliable predictor of high counts between two HEIs than organizational size. This is consistent with descriptive statistics showing that a very high number of webpages (in respect to organizational size) does not translate into high total degree.

Finally, the full model classifies 93% of the cases in the same category as the size model, showing that adding information on the websize (normalized) to the model might somewhat influence the predicted number of counts, but it is not changed drastically. The predictive ability of the two models is not significantly different.

5 Discussion and conclusions

Before coming to our conclusions, it is relevant to highlight the limitations related to the sample considered. Namely, our sample provides a good coverage of doctorate-awarding universities in Europe (about 50% of those included in the EUMIDA restricted set, comprising more than 70% of staff and students), but does not include the large number of non-doctorate awarding HEIs. We however provided in section 3.2 preliminary evidence that some of our results (those concerned with the size of the web

domains and total degree) are likely to be valid for all HEIs in Europe. Importantly, the sample considered in the paper accounts for most of the connectivity in European higher education. Accordingly, our results hold for the most central and more connected part of European higher education, while we cannot exclude that for peripheral HEIs, the mechanisms driving connectivity are slightly different.

With these limitations in mind, the paper provides novel results concerning the factors influencing the size of the HEI websites, respectively on their impact on the HEI connectivity on the web. On the one hand, the number of webpages of HEIs turns out to be largely determined by organizational size, as well as to a lesser extent by the volume of educational and research activities. We could however identify a small numbers of HEIs whose websites are far larger than expected from their size. Differences in the web presence of HEIs across European countries have clearly diminished in the last years. The academic web seems then to have reached a mature stage.

On the other hand, our results show that the impact of the webdomain size on the number of weblinks between two HEIs remains relatively limited: having more webpages (relative to size) increases somewhat the number of links when two HEIs are already connected – and hence increases also total degree -, but does not have a significant impact on the likelihood two HEIs being connected. As a matter of fact, organizational size is a more reliable predictor of the number of weblinks between two HEIs than the size of their website, since having a large number of webpages (as compared to size) does not translate into stronger connectivity. This is important for studies using weblinks for two reasons: first, not always information on the size of the respective webdomains is available and, second, interlinking patterns between HEIs are then reasonably robust against changes in the amount of information published and thus are more likely to be stable across time. Of course, we cannot fully exclude that other characteristics of websites, like how links are integrated into webpages, impact on connectivity; however, our results suggest that these might influence counts in individual cases, but are not very likely to generate new connections and fundamentally modify the network structure.

We explain the observed patterns by the tendency to standardization in the structure and management of academic websites and by the fact that weblinks represent markers of social relationships between HEIs rather than between documents published on the Web. On the one hand, we hold anedoctal evidence that the structure of academic websites tends to mirror organizational structure and that standards are diffusing concerning the number of webpages for specific activities and, especially, personal webpages (supported by the diffusion of CMS). This does not exclude that some HEIs invest in publishing large amounts of contents in educational sites and repositories of research products, but this seems not to be a widespread phenomenon.

On the other hand, results concerning weblinks are more consistent with a model where hyperlinks are markers of underlying social relationships between HEIs, than with a model where they are related to the specific content of the linked webpages. The academic web then presents a different structure than other web domains or co-authorship networks, with their highly-skewed power-law distribution of connectivity; on the academic web, almost all HEIs enjoy of some level of visibility, while connectivity is not concentrated in a few HEIs (even if there are HEIs more connected than the average).

In broader terms, our results support previous remarks that weblinks are a fundamentally different phenomena than citations or co-authorships of academic papers and, accordingly, they are not likely to replace citations as reputational measures for HEIs (Thelwall 2012b). While this does not necessarily apply to all academic websites – journal websites are likely to display more content-related patterns - it is definitely the case for organizational websites of HEIs. At the same time, they emphasize the potential of weblinks in order to study organizational networks between HEIs, as associated to the underlying

distribution of resources and status, as well as to the institutional and geographic structure of higher education systems.

Finally, our research opens a few relevant areas of future investigation. First, there is limited information available on the web policies and organization of HEI websites, as most available studies were concerned with issues of layout and usability to external audiences like students. In-depth analyses through case studies, as well as providing detailed counts of pages by subdomains, would be useful to shed further light on this issue. Relatedly, it would be important to provide more systematic analyses of repositories published on the Web by HEIs and to the extent they account for a large share of webpages and of connectivity for individual HEIs. Investigation of this aspect is highly relevant since we cannot exclude that in the future repositories account for a larger share of HEI websites.

Finally, it would be highly relevant to investigate whether links between HEI websites and websites of other organizations – public authorities, NGOs, companies – display the same patterns; should this be the case, they would become an important measure to investigate HEI relationships with society and economy, a domain where there is clearly the need of additional data.

6 References

Adamic, L. A. & Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science, 287(5461)*, 2115-2115.

Aguillo, I. F., Ortega, J. L., Fernández, M. & Utrilla, A. M. (2010). Indicators for a webometric ranking of open access repositories. *Scientometrics, 82(3)*, 477-486.

Aguillo, I., Ortega, J. L. & Fernández, M. (2008). Webometric Ranking of World Universities: Introduction, Methodology, and Future Developments. *Higher Education in Europe, 33(2)*, 233-244.

Almind, T. C. & Ingwersen, P. (1997). Infometric analyses on the world wide web: methodological approaches to 'webometrics'. *Journal of Documentation, 53(4)*, 404-426.

Astani, M. (2013). A Decade of Changes in University Website Design. *Issues in Information Systems, 14(1)*, 189-196.

Barabási, A., Albert, R. & Jeong, H. (2000). Scale-free characteristics of random networks: the topology of the world-wide web. *Physica A: Statistical Mechanics and its Applications, 281(1)*, 69-77.

Bar-Ilan, J. (2005). What do we know about links and linking? A framework for studying links in academic environments. *Information Processing & Management, 41(3)*, 973-986.

Bar-Ilan, J. (2009). Infometrics at the beginning of the 21st century - A review. *Journal of Infometrics, 2(1)*, 1-52.

Callahan, E. (2005). Cultural similarities and differences in the design of university web sites. *Journal of Computer-Mediated Communication, 11(1)*, 239-273.

Callahan, E. & Herring, S. C. (2012). Language Choice on University Websites: Longitudinal Trends. *International Journal of Communication, 6*, 322-355.

Capocci, A., Servedio, V. D., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S. & Caldarelli, G. (2006). Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical Review E, 74(3)*, 036116.

Cattani, G., Ferriani, S., Negro, G. & Perretti, F. (2008). The Structure of Consensus: Network Ties, Legitimation, and Exit Rates of U.S. Feature Film Producer Organizations. *Administrative Science Quarterly, 53(1)*, 145-182.

Daraganova, G., Pattison, P., Koskinen, J., Mitchell, B., Bill, A., Watts, M. & Baum, S. (2012). Networks and geography: Modelling comunity network structures as the outcome of both spatial and network processes. *Social Networks, 34(1)*, 6-17.

De Stefano, D., Fuccella, V., Vitale, M. P. & Zaccarin, S. (2013). The use of different data sources in the analysis of co-authorship networks and scientific performance. *Social Networks, 35(3)*, 370-381.

Gaudeul, A. & Giannetti, C. (2013). The role of reciprocation in social network formation, with an application to LiveJournal. *Social Networks, 35(3)*, 317-330.

Glänzel, W. & Schubert, A. (2005). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel & U. Schmoch(Eds.) *Handbook of Quantitative Science and Technology Research* (pp. 257-276). Dordrecht: Kluwer Academic Publications.

Gonzalez-Bailon, S. (2009). Opening the black box of link formation: Social factors underlying the structure of the Web. *Social Networks, 31(4)*, 271-280.

Heller-Schuh, B., Barber, M., Henriques, L., Paier, M., Pontikakis, D., Scherngell, T., Veltri, G. A. & Weber, M. (2011). *Analysis of Networks in European Framework Programmes (1984-2006)* Luxembourg: Publications Office of the European Union.

Ijiri, Y. & Simon, H. A. (1964). Business firm growth and size. *The American Economic Review, (54(2))*, 77-89.

Katz, J. S. & Cothey, V. (2006). Web indicators for complex innovation systems. *Research Evaluation, 15(2)*, 85-95.

Kretschmer, H., Kretschmer, U. & Kretschmer, T. (2007). Reflection of co-authorship networks in the Web: Web hyperlinks versus Web visibility rates. *Scientometrics, 70(2)*, 519–540.

Lee, M. & Park, H. W. (2012). Exploring the web visibility of world-class universities. *Scientometrics, 90*, 201-218.

Lepori, B., Barberio, V., Seeber, M. & Aguillo, I. (2013). Core–periphery structures in national higher education systems. A cross-country analysis using interlinking data. *Journal of Infometrics, 7(3)*, 622–634.

Lepori, B. & Bonaccorsi, A. (2013). Towards an European census of higher education institutions. Design, methodological and comparability issues. *Minerva, 51(3)*, 271-293.

Más-Bleda, A. & Aguillo, I. F. (2013). Can a personal website be useful as an information source to assess individual scientists? The case of European highly cited researchers. *Scientometrics, 96(1)*, 51-67.

Ortega, J. L., Aguillo, I., Cothey, V. & Scharnhorst, A. (2008). Maps of the academic web in the European Higher Education Area - an exploration of visual web indicators. *Scientometrics, 74(2)*, 295-308.

Owen-Smith, J. & Powell, W. W. (2008). Networks and institutions. In R. Greenwood, C. Oliver, K. Shalin & R. Suddaby(Eds.) *The SAGE Handbook of Organizational Institutionalism* (pp. 594-621). London.

Payne, N. & Thelwall, M. (2007). A longitudinal study of academic webs: Growth and stabilization. *Scientometrics, 71(3)*, 523-539.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J. & Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences, 99(8)*, 5207-5211.

Peterson, K. (2013). Academic web site design and academic templates: Where does the library fit in? *Information Technology and Libraries, 25(4)*, 217-221.

Rivera, M. T., Soderstrom, S. B. & Uzzi, B. (2010). Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms. *Annual Review of Sociology, 36*, 91-115.

Seeber, M., Lepori, B., Lomi, A., Aguillo, I. & Barberio, V. (2012). Factors affecting weblink connections between European higher education institutions. *Journal of Infometrics, 6(3)*, 435-447.

Thelwall, M. (2002a). A research and institutional size based model for national university web site interlinking. *Journal of Documentation, 58(6)*, 683-694.

Thelwall, M. (2002b). Evidence for the existence of geographic trends in university web site interlinking. *Journal of Documentation, 58(5)*, 563-574.

Thelwall, M. (2012a). A history of webometrics. *Bulletin of the American Society for Information Science and Technology, 38(6)*, 18-23.

Thelwall, M. (2012b). Journal impact evaluation: a webometric perspective. *Scientometrics, 92(2)*, 429-441.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57(1)*, 60-68.

Thelwall, M., Binns, R., Harries, G., Page-Kennedy, T., Price, L. & Wilkinson, D. (2002). European Union associated university websites. *Scientometrics, 53(1)*, 95-111.

Thelwall, M. & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organizations. *Journal of the American Society for Information Science and Technology, 62(8)*, 1488-1497.

Thelwall, M., Tang, R. & Price, E. (2003). Linguistic patterns of academic web use in Western Europe. *Scientometrics, 56(3)*, 417-432.

Vaughan, L. & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing & Management, 41(2)*, 347-359.

Vaughan, L., Kipp, M. & Gao, Y. (2007). Why are Websites co-linked? The case of Canadian Universities. *Scientometrics, 72(1)*, 81-92.

Wagner, C. S. & Leydesdorff, L. (2005). Network Structure, Self Organization, and the Growth of International Collaboration in Science. *Research Policy, 34 (10)*, 1608-1618.

Wilkinson, D., Harries, G., Thelwall, M. & Price, L. (2003). Motivations for academic web site interlinking: evidence for the Web as a novel source of information on informal scholarly communication. *Journal of Information Science, 29(1)*, 49-56.