Journal of Informetrics 6 (2012) 435–447

Contents lists available at SciVerse ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Factors affecting web links between European higher education institutions

Marco Seeber[a,*], Benedetto Lepori[b], Alessandro Lomi[b], Isidro Aguillo[c], Vitaliano Barberio[d]

[a] *Centre for Organizational Research, University of Lugano, Via Lambertenghi, 6900 Lugano, Switzerland*
[b] *Centre for Organizational Research, University of Lugano, Switzerland*
[c] *Cybermetrics Lab, CSIC, Madrid, Spain*
[d] *Department of Public Management, WU University, Vienna, Austria*

ARTICLE INFO

Article history:
Received 22 November 2011
Received in revised form 2 March 2012
Accepted 5 March 2012

Keywords:
Webometrics
Web links
Antecedents of relationships

ABSTRACT

We examine the extent to which the presence and number of web links between higher education institutions can be predicted from a set of structural factors like country, subject mix, physical distance, academic reputation, and size. We combine two datasets on a large sample of European higher education institutions (HEIs) containing information on inter-university web links, and organizational characteristics, respectively. Descriptive and inferential analyses provide strong support for our hypotheses: we identify factors predicting the connectivity between HEIs, and the number of web links existing between them. We conclude that, while the presence of a web link cannot be directly related to its underlying motivation and the type of relationship between HEIs, patterns of network ties between HEIs present interesting statistical properties which reveal new insights on the function and structure of the inter organizational networks in which HEIs are embedded.

1. Introduction

During the last decade increasing attention has been dedicated to the study of connections between higher education institutions (HEIs) through their web domains (Bar-Ilan, 2009). Web sites are important coordination devices that may be used to support a wide range of inter organizational communications (Thelwall & Zuccala, 2008). A number of studies are available that investigate the motivation behind their creation (Bar-Ilan, 2004; Vaughan, Kipp, & Gao, 2007), as well as the structure of interlinking within and between European countries (Ortega, Aguillo, Cothey, & Scharnhorst, 2008; Thelwall, 2002b).

Against this background, in the paper we want to identify the major factors influencing the probability of the creation of web links between two HEIs. Potential antecedents include institutional factors defined, for example, in terms of national and linguistic boundaries, the distance between HEIs and organizational factors such as size and research quality. While previous research agrees that these factors affect the presence and number of web links (Thelwall, 2002a), their relative strength has never been investigated on a sample large enough to draw robust conclusions and to generalize results beyond national situations.

The analysis is based on a sample of 1181 HEIs in 28 European countries obtained by matching interlinking data provided by the Cybermetrics lab (Ortega et al., 2008) with structural characteristics of HEIs derived from the EUMIDA dataset (Bonaccorsi et al., 2010). The matching of the two datasets represents and important innovation, which allows a better

* Corresponding author. Tel.: +41 58 666 4624; fax: +41 58 666 4624.
E-mail addresses: marco.seeber@usi.ch (M. Seeber), benedetto.lepori@usi.ch (B. Lepori), alessandro.lomi@usi.ch (A. Lomi), Isidro.aguillo@cchs.csic.es (I. Aguillo), vitaliano.barberio@gmail.com (V. Barberio).

1751-1577/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2012.03.001

understanding of the relationship between web links, the characteristics of individual HEIs and their relative position in the institutional and physical space of European higher education.

We organize the paper as follows. In the next section we introduce our approach to modeling weblinks, while in Section 3 we present the dataset and the measures of antecedents. In Section 4 we provide a descriptive analysis of web links, while in Section 5 we report the results of inferential analyses of the antecedents of network ties between HEIs and of their strength. We conclude the paper by discussing the methodological and substantive implications of these results for the study of network relations between HEIs.

2. Background and theoretical framework

The conceptual framework of this paper can be outlined as follows: HEIs are connected through a web of relationships related to a diverse set of activities and motivations. Social network theory predicts that their presence and strength is influenced by a set of factors through *assortative* and *proximity* mechanisms (Rivera, Soderstrom, & Uzzi, 2010). Previous research shows that Web links provide a synthetic indicator for relationships and, accordingly, we expect their presence to depend on these factors.

The relevance of this topic goes beyond the technical issue of modeling hyperlinks. An increasing body of literature in the sociology and economics of organizations demonstrates that social relationships are a key determinant of organizational behavior and performance, as these provide key resources and opportunities to individual organizations (Brass, Galaskiewicz, Greve, & Tsai, 2004; Granovetter, 1985), while, at the field level, network structures frame economic exchange and market structure (White, 2002). Accordingly, inter-organizational networks, rather than individual organizations, are increasingly considered as the locus of organizational learning, innovation and performance (Powell, Koput, & Smith-Doerr, 1996).

These arguments are well-known in sociology and economics of science, which widely investigated the collaborative dimension of scientific and technological activities, i.e. through studies of co-authorship of academic papers (Glänzel & Schubert, 2005), as well as of patent networks (Breschi & Lissoni, 2004). Networks in science are not limited to *collaborations*, but also involve other forms of *relationships* like indirect exchange of knowledge, information, status and reputation – elements that are at the heart of science as a collective enterprise (Merton, 1968).

While network studies are relatively widespread at the individual and research group level (Braun & Glänzel, 1996; Glänzel, 2001; Schubert & Braun, 1990; Wuchty, Jones, & Uzzi, 2007), few studies have until now investigated the relational structure between higher education institutions as a whole, partially because of data limitations, but also of an understanding of HEIs as a mere container of disciplines (Clark, 1983; Weick, 1976) and thus that the institutional level was not particularly relevant.

However, two related trends – namely the transformation of HEIs towards complete organizations with a strategic function (de Boer, Enders, & Leisyte, 2007) and an increasing role of market coordination in the system's governance (Deiaco, Holmén, & McKelvey, 2010) – heighten the importance of a thorough analysis of relational structures to understand the evolution of higher education, as well as behavior and performance of individual HEIs.

2.1. Weblink analysis: a review of literature

In the last decade, web links have received an increasing attention in studies of inter-institutional relationships (Bar-Ilan, 2009), as possible indicators of relationships between higher education institutions together with data on co-authorships (Jones, Wuchty, & Uzzi, 2008) and participation to European Framework Programs (Heller-Schuh et al., 2011). These works mainly deal with two issues: (i) understanding the motivations behind the creation of weblinks, and (ii) to modeling weblink as a function of organizational characteristics.

a) *Motivation studies*. Studying the motivations of interlinking is critical as motivation cannot be directly related to specific types of relationships, unlike in the case of co-authorships and project collaborations. Different classification schemes have been developed, considering characteristics of the source page, of the link (i.e. content) and of the target page (Bar-Ilan, 2004, 2005). Micro-level studies show that motivations for weblinks are broader than research collaboration, but are largely related to the main activities of HEIs (Thelwall, 2002b). Thus, about 90% of links between UK universities are created for scholarly reasons, whereas only a minor share (1%) is equivalent to online citations (Wilkinson, Harries, Thelwall, & Price, 2003). Expectedly, hyperlinks do not reflect collaboration structures derived by bibliometric data (Kretschmer, Kretschmer, & Kretschmer, 2007). Remarkable differences also emerge in the use of links within the same discipline (Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004).

Motivations for interlinking include signaling the institutional space to which the HEI belongs (links to all similar departments in a country), referring to useful information in the same geographical area (links to services in universities nearby), referring to educational materials, to academic cooperation partners, signaling deference to the institution or groups considered as the best in a specific area (Bar-Ilan, 2004). Working on a sample of Israeli universities webpages, Bar-Ilan found that the main motivations for links creation were professional and work related (32%), research oriented (28%) and informative (14%) (Bar-Ilan, 2005). Moreover, she showed that weblinks display a hub structure, with most of the links emanating from list pages and pointing towards information sources at different places. As lists tend to be based on proximity relationships – belonging to same geographical and institutional space – this hints to a specific mechanism through which these mechanisms influence interlinking patterns.

b) *Modeling and empirical testing*. A second line of studies analyzed the impact of antecedents on the number of weblinks between HEIs. In the UK case, the best fit for the number of weblinks between pairs of universities was provided by the product of the size and research quality (average RAE scores) of the sending and the receiving institution (Thelwall, 2002a). The relationships between interlinking and research quality is more nuanced in other studies, which points to the importance of research productivity, as universities producing more research also produce more web pages but with a similar average online impact (Almind & Ingwersen, 1997; Thelwall & Harries, 2004). Other works suggest that, despite the diffusion of electronic communication, the distance between HEIs still has an impact on number of weblinks, even if interlinking remains strong among top class universities (Thelwall, 2002a,b).

Studies at the European level were largely content to describe network structures, without attempting to model weblinks. Thus, Ortega et al. (2008) found that the European-level interlinking patterns are set up by the aggregation of national networks, whereby the German and British ones are dominant. A university is first linked to others within its country and then to other national networks. Some university web sites act as gatekeepers between the national and the European networks (Ortega et al., 2008). Some web sites' and university characteristics emerged to affect patterns of international connections. In a study of 16 European countries, English accounted for half of the international linking pages and universities tended to link most to countries with shared language or geographically close (Thelwall, Tang, & Price, 2003). International linking is also associated with country size and there is some clustering amongst countries (Heimeriks & Van den Besselaar, 2006).

Finally, a recent study on over 400 European life science research group web sites made a systematic attempt at modeling weblinks at the international level through a multivariate analysis, showing that research group size and web presence were important for attracting web links, although not research productivity (Barjak & Thelwall, 2008).

This review of previous research leads to a few relevant remarks. First, there is evidence that the set of motivations for establishing a weblink is much broader than the existence of a research collaboration. As a consequence, modeling weblinks needs to account for the specific *mechanisms* generating relations between organizations. Second, empirical studies identify antecedents – including size, research quality and productivity, geographical distance, country and language – and display quantitative relationships between antecedents and number of weblinks. However, the generalizability of these results is limited as the sample on which they are based include only one or few countries. Third, data quality problems are relevant and, accordingly, both sample size and statistical techniques need to be robust. In particular the non-normal distribution of weblinks confronts analysts with the problem of dealing with outliers. This is only in part a statistical problem as the presence of institutions attracting or generating high number of links reveals behavior that need theoretical or empirical explanation (particularly against biases generated by outliers).

2.2. *Antecedents of web link connections*

The literature on social networks identifies assortativity and proximity as the main relational mechanisms influencing the existence, strength and stability of dyadic ties between two organizations (Rivera et al., 2010).

Assortativity links the creation, persistence and dissolution of social relationships to the similarity, compatibility and complementarity in the actors' attributes; in some cases actors with similar attributes are more likely to develop a relationship, in other cases it is complementary attributes that favor relationships, for instance when a variety of competences is needed. In higher education institutions, *assortative (dissortative) mechanisms* refer to similarity (difference) in the attributes of individual HEIs, like mission (research vs. teaching oriented), the range of subjects covered, legal status, language, research quality.

Proximity mechanisms affect interactions in social, institutional and geographical spaces because the likelihood of establishing relationships increases when two organizations share the same space. The simplest case is represented by geographical or physical proximity, which has been shown to affect the formation of social and inter-organizational ties. In the case of higher education institutions, relevant dimensions of proximity are related to geography, belonging to the same country, region and to status and reputation effects.

Distinguishing empirically between assortative and proximity mechanisms is not always straightforward as individual attributes and proximity are likely to be correlated; accordingly, sufficiently large samples and the use of appropriate statistical techniques are required to ascertain their relative importance.

a) The *mission* is an important element defining the space for action and, thus, for relationships. HEIs with the same mission are more likely to compete for education and to share similar research goals and interests. A key issue in this respect is research orientation; as research requires intense exchange and communication, research oriented HEIs are expected to be more interconnected than those focused on teaching.

b) The similarity concerning *subject domains* should increase the number of web links: HEIs are more likely to cooperate and compete when their subject mix is similar as research and educational activities are organized by subject and cross-disciplinary relationships are relatively rare.

c) Sharing the same *language* eases communication and, in turn, collaboration. Thus, two HEIs with the same language will be more likely to establish some kind of relationships. Moreover, a positive bias in favor of *English* speaking HEIs would be expected, because English is the most widespread language in Science.
d) The *legal status* of an HEI constraints and influences its behavior and pattern of relationships by means of regulation, market forces and perceptions of the actors. Thus, HEIs with the same legal status are expected to be more interconnected.
e) As highlighted by some studies the *research productivity* affects the number of links received at the group level, because research works are a relevant target of links, but the impact at the institutional level is more uncertain as research related links only sum up a small share of the total.
f) As HEIs are largely institutionalized organizations, whose mission, rules and resources depend from the State, belonging to the same country is likely to generate more intense relationships: HEIs in the same country tend to be more similar and share elements of cultural, historical, and institutional context and conveners, while government policies may be important to stimulate the emergence of networks (Doz, Oik, & Ring, 2000). This is confirmed by all previous studies of weblinks.
g) The likelihood of being linked and their strength is expected to increase with *geographic proximity* between two HEIs, as physical distance is a relevant fact in almost all network studies and this applies at least for some of the HEI relationships (like educational cooperation and sharing services), possibly to a lesser extent for research where relationships tend to be increasingly geographically spread (Wagner, 2008),
h) *Research quality* is expected to impact on relationships as a tendency of HEIs to cooperate when their level of quality is similar is expected. Further, reputation – which in higher education is closely related to research quality – is one of the main factors structuring social networks (Burris, 2004). Existing studies display a strong tendency to social stratification in multi-university collaborations (Evans, Lambiotte, & Panzarasa, 2011; Jones et al., 2008), with a core of top-ranked institutions forming exclusive relationships with one another (Opsahl, Colizza, Panzarasa, & Ramasco, 2008). Accordingly, the reputation of the receiving and sending HEIs, as well as their relative position, are expected to influence numbers of web links.
i) Finally, the size of sender and receiver HEIs are expected to influence the number of weblinks, as the volume of activities, as well as the number of webpages, are expected to be roughly proportional to size.

2.3. Modeling weblinks

To model weblinks, we consider counts of weblinks between two webdomains as the aggregated outcome of a probabilistic process driving the chance that two individual webpages are linked together. Moreover, we assume that the probability of individual links is a function of the antecedents introduced in the previous section. Accordingly, we rely on techniques used for modeling count data for series of non-negative integers possibly including many zeros (Cameron & Trivedi, 1998).

If individual events are independent and their number is sufficiently large, the resulting probability distribution for the counts follows a Poisson distribution:

$$Pr(N; m) = \frac{e^{-m} m^N}{N!}, \quad \text{where} \quad N = \text{number of counts}$$

Then, the expected number of counts $E(N) = m$. As m increases, this distribution approximates a normal distribution, but if m is small it displays a right-skewed distribution.

If m is a function of a set of antecedents, $m = m\ (x)$, a Poisson regression with the form $m(x) = e^{\beta x}$ can be used and the coefficients can be estimated through maximum likelihood. In that case, the expected number of counts is $E(x) = m(x) = e^{\beta x}$ while their variance is also $V(E) = m(x)$. Thus, unlike linear regressions, the Poisson regression model does not assume that observations are normally distributed around the conditional mean, while there is heteroskedasticity (i.e. the variance is increasing with the conditional mean), two well-known characteristics of weblinks statistics.

Descriptive statistics show that weblink data are characterized by overdispersion (i.e., the tendency of the variance to increase faster than the mean), we use a *negative binomial regression* which includes a parameter to model overdispersion. Further, since the number of null dyads (dyads with no links) is very high (88% of the sample), we use a *hurdle* negative binomial, which specifies a separate model for predicting zeros – the underlying assumptions being that factors explaining zeros might be different from those explaining counts (Mullahy, 1986).

The interpretation of the regression results differs from ordinary regressions. First, the model provides expected count values of weblinks $E(x)$, but it is not assumed that the distribution of observed counts is normal; accordingly, there is no straightforward interpretation in terms of distribution of observations around the expected value and usual fit measures based on these assumptions (like R^2) cannot be used. Instead, the fit of the model may be evaluated by the percentage of observed counts correctly predicted.

Second, binomial regression coefficients are exponential and multiplicative: if the coefficient for an antecedent is β, then the *percentage change* in the expected number of counts for unit a change in the antecedent is e^{β}. Changes in different antecedents have a multiplicative impact on expected number of weblinks; this corresponds to previous works showing that numbers of weblinks can be modeled from the product of sender's and receiver's size and quality (Thelwall, 2002a).

3. Data sources and methods

A major element of innovation in our work derives from matching the number of weblinks connections between HEIs with a data set containing information on their individual characteristics. Previous studies either worked on a single national context (Bar-Ilan, 2004; Vaughan et al., 2007), or relied on webometrics data only (Ortega et al., 2008).

Interlinking data. The interlinking data were obtained from commercial public search engines following the methodology described in Aguillo, Granadino, Ortega, and Prieto (2006). Two mirrors of the *Yahoo Search!* database were used, the Spanish and the British ones, to avoid collection problems derived from restrictions in the limited bandwidth available or from errors in the automatic scripts used for extracting the data. If the results for the same request were not identical, then the maximum value of the two was used. The collection took place in January 2011.

The web domains for 1337 European universities in the EUMIDA database were identified. When an institution had more than one central domain, the one receiving the larger number of external in-links was chosen. This means that in some cases the present web domain is not used because it has less link visibility than the older one. As the interlink pairs are directional, the requests were performed on the full matrix of 1337 × 1337 items.

Organizational data have been derived from the EUMIDA (European Micro Data) dataset, which includes information on HEIs in 28 European HE systems (European Union members plus Norway and Switzerland, France not included). EUMIDA consists of two samples, one of 2457 HEIs and the other, more detailed, of 1378 'research active' HEIs (Bonaccorsi et al., 2010); these include 850 doctorate-awarding institutions and comprehend 82% of the total number of students at tertiary level. The data mostly refer to the year 2008, and include: identifying information (name, category, foundation year, highest degree granted), expenditure and income, academic and non-academic staff, bachelor and doctoral students per total number and by field, patents and spin-off, degree awarded by national/international origin and by field.

Each web domain has been related to the corresponding HEI in EUMIDA in order to allow the matching of the two datasets. A matrix 1182*1182 was created with HEIs that had all the organizational data required for the analysis.

Outliers. We investigated whether some of the strongest connections could be generated by technical artifacts, focusing on connections where an HEI receives more than 10,000 web links from another HEI and this represents over 40% of the total web links that it receives; this procedure identified 32 'suspect' connections. Some of them are due to technical factors: thus, the 'soton.ac.uk' domain is often the strongest linker because 'eprints.ecs.soton.ac.uk' is the home of the software for e-prints repositories and inflates artificially the web links flows to other institutions. This domain has been removed from the sample. Also, the connection 'uni-trier.de' to 'rwth-aachen.de' (318,000 links) is abnormally high, due to the "DBLP-Digital Bibliography Library Project" repository, one of the largest database about computer science that has two mirrors. We removed the links related to DBLP (312,000). The remaining connections originate when two universities are located in the same city or even in the same campus, when they share a common repository or they are scheduled to be merged soon.

Previous work has shown that the use of alternative document models (ADM) other than individual webpages is a solution to this problem and improves the fit of weblinks modeling (Thelwall, 2002a). Unfortunately, as our data have been collected from commercial search engines, ADM cannot be applied. However, the issue of outliers is less relevant for our analysis for two reasons: first, we are focusing on the large-scale statistical patterns of interlinking, while outliers represent an exceedingly small share of the dyads; second, as binomial regressions are based on maximum likelihood rather than on least squares, they are less sensitive to outliers than ordinary regressions and thus it is unlikely that results are biased by their presence.

Antecedents. Various matrices have been created to test hypotheses on the antecedents of relationships. In the matrices, *each cell contains a value or code representing the relationship between the HEI sending the link (sender) and the HEI receiving the link (receiver)*. The first matrix contains the number of web links sent from one HEI to another (the values of the diagonal have not been considered), while matrices of independent variables contain a value or code representing the relationship between the sending and receiving HEI (Table 1). *Each dyad is treated as independent, thus we do not address the issue of network structure*.

The indicator of *distance* measures the distance in kilometers between two HEIs. Each web domain corresponds to an IP, which has been related to the latitude and longitude coordinates used to compute the distances. Manual data cleaning identified the cases when IP did not correctly locate the university (about 5% of the sample).

The indicator for the *mission* points out whether the two HEIs are both PhD awarding or not. This indicator alone can hardly comprehend the complexity of a university mission, though the possibility to award the PhD certificates is a major element differentiating different types of HEIs.

Discipline similarity expresses the extent to which the HEIs have a similar subject mix. The share of students enrolled by each of the nine subject domains of educational statistics has been computed; the subject overlap between HEI "*x*" and HEI "*y*" is given by the following formula:

$$\text{Subject Overlap} = \sum_{i=1\to 9} \text{MIN}\,(x_i, y_i)$$

where x_i and y_i represent the share of students in a given discipline *i*. The indicator ranges from 1, when the two HEIs have the same disciplinary profile, to 0, when they have a completely different profile. A limitation of this measure is that it does

Table 1
Antecedents, hypothesis, data and variables.

Antecedents	Hypotheses	Data	Independent variables
Context	HEIs in the same context cooperate and compete more, because they tend to be more similar and draw resources from the same environment	Country code	1 if the HEIs are in the same country, 0 otherwise
		NUTS 2[a] region code	1 if the HEIs are in the same region, 0 otherwise.
Geographic proximity	Geographically closer universities are more interconnected because proximity eases communication	Distance	Kilometric distance between the sending and the receiving HEIs
Mission	Universities awarding PhD will be more interconnected	PhD awarding	1 if both HEIs award doctorate degree, 0 otherwise
Subject similarity	HEIs with a similar discipline specialization will be more inter connected	Specialization by discipline (students)	Index of similarity in the discipline profile, from 0, no similarity, to 1, maximum similarity
Research productivity	Universities with high research productivity per unit of academic staff send and receive more links	Leiden ranking	Average "P" per unit of academic staff
Size	Larger HEIs will receive more links	Academic staff	No. of academic staff of the receiver HEI
	Larger HEIs will send more links	Academic staff	No. of academic staff of the sending HEI
Language	HEIs with a common language will cooperate more	Language code	1 if the HEIs share the same language, 0 otherwise
	English speaking HEIs will be more linked	Language code	1 if the receiving HEI is located in an English-speaking country, 0 otherwise
Legal status	HEIs with a similar status are most likely to compete and cooperate	Public, private, government dependent	1 if the HEIs have the same legal status, 0 otherwise
Research quality	Universities with high research quality are more reputed and tend to send and receive more links	Leiden ranking	Weighted IF of the sending and target HEIs

[a] NUTS stands for '*nomenclature d'unités territoriales statistiques*; they are geocode standard for subdivisions of countries developed and regulated by the European Union. There are three levels. NUTS 2 are 271 and each region, with some exceptions, sums up between 800 thousands and 3 million inhabitants.

not take into account that some disciplines may be more prone to collaborate with each other than with others. Moreover, organizations may teach in areas where they do not actually research or vice versa.

Productivity indicator measures the average productivity per researcher and it is computed by the ration between the number of publications and the number of academic staff. The indicator of *Quality* measures the average quality of the publications and it given by their field-normalized average impact. The source of these data is the Leiden ranking (2008).[1] As data on productivity and quality were only available for those HEIs comprised in the ranking, the impact of these factors have been explored with a specific test on this sample. We chose this ranking for three main reasons. First, it is rather stable to variation from one year to another, since it is computed by considering the productivity of the last eight years (2000–2007). Second, among the most renown rankings, it has the largest sample of European HEIs (250). Finally, it provides both the number of publications (*P*) and the university's field-normalized average impact.

EUMIDA identifies three categories of *legal status* across the considered countries: public, private, government dependent. The categorical variable indicates whether the HEIs have the same legal status or not.

4. Descriptive statistics

The sample includes 1181 HEIs: 731 of them award PhD certificates, and 182 are in the Leiden ranking, 937 are public, 154 private and 90 are government-dependent private, i.e. private institutions that receive most of their funding from the government. The largest country in terms of representation in the sample are Germany (292), UK (145), Poland (85) and Italy (80). German is the predominant language (379) followed by English (166 – Malta not considered) (Table 2).

Dyadic QAP correlations[2] (with UCINET) shows that most variables are weakly correlated, with the exception of *language* and *country* (0.815, 5000 permutations, *P*-value = 0), since in Europe, country and language often coincide.

The distribution of web links is strongly right skewed (Table 3) – a result consistent with previous studies demonstrating that the frequency distribution of links follows power laws (Broder et al., 2000). Total number of cells is 1,393,580.

88% of the dyads are not active (i.e. null). This mean that overall, the European HEIs are largely (directly) disconnected. This is not particularly surprising, since many HEIs in the sample are small, and with a teaching and local orientation. Nine percent of the matrix comprises values between 1 and 10. Less than 3% of the connections is composed by dyads above 10 links but they represent about 90% of the total number of links sent.

Metaphorically speaking, there are large areas of 'dark' and a small area of intense "light". It follows that a major task is to identify where connections are located.

[1] http://www.cwts.nl/ranking/LeidenRankingWebSite.html.

[2] QAP is used as matrix correlations are flawed.

Table 2
Antecedents' main statistics.

Variable	S.D.	Min	1° Quartile	Mean	Median	3° Quartile	Max
Distance	844	0	800	1435	1329	1950	6413
Subject	0.27	0	0.09	0.34	0.32	0.56	1
Size	870	4	111	680	320	908	6571

Variable	Dummy = 1	Dummy = 0
Country	144,556	1,249,024
PhD	533,630	859,950
NUTS regions	10,010	1,383,570
English	195,880	1,197,700
Legal Status	908,604	484,976
Language	204,159	1,189,421

Table 3
Descriptive statistics on the matrix of web links.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	2918	0.000	50,100
Standard deviation: 118.07			Gini coefficient: 0.982		

	Ranges in the number of links							
	0	1–10	11–50	51–250	251–1000	1001–10,000	>10,000	Total
No. dyads	1,226,453	128,328	28,910	8050	1,445	366	28	1,393,580
Total links	0	401,396	623,092	839,794	647,612	976,630	581,600	4,070,124
Share of dyads	88%	9%	2%	0.6%	0.1%	0.03%	0.002%	
Share of links	0%	10%	15%	21%	16%	24%	14%	

The antecedents considered have a strong impact both on the share of active connections (Table 4), as well as on the ratio between the number of strong (>10 web links) and weak connections (1–9 web links). For example, two HEIs are in the same country have a probability of being linked of 32% against only 10% if they are in different countries. When both HEIs award PhD degrees and are located in the same country, then 59% of the dyads are active; conversely, 97% of the dyads between non PhD awarding HEIs in different countries are non-active.

In summary, descriptive analysis shows that the distribution of active connections departs from randomness and that most antecedents are relevant to identify areas of the matrix with active and intense connections.

5. Testing antecedents of web links

Table 5 presents the results of negative binomial regressions on the full sample for three models. The country model is superior to the null model, but the complete model is largely superior to both. The third model represents the optimal balance between fit and the number of variables included, as adding English, Legal Status, Language, or cross-terms does not improve the statistical performance meaningfully. The function 'hurdle' separately predicts estimates for the zero values and for the positive values. This is particularly valuable because our first goal is to distinguish non-active from the active connections.

All the parameters are significant and with the expected sign. Regression significance can be evaluated by comparing predicted and actual values on several intervals of web links and thus looking to the ability of the model to predict the observed values (Table 6).

The first two intervals distinguish active from non-active connections, representing 12% and 88% of the connections each. Thus, the model identifies 85% of the zeros (sensitivity) and, when it predicts zero, it is correct in 96% of the cases (positive predictive value). The performance is also good in terms of detecting the active connections (75%, specificity); when the model predicts an active value, it is correct in 41% of the cases. Overall, the capability of the model to detect active and non-active connections appears rather good.

The second test focuses on the identification of the 'strong' connections, i.e., those above 10 links, representing less than 3% of the dyads but summing up 90% of the links sent. On the one hand, the model detects 97% of the non-strong connections and, in that case, it is correct in the 99% of the cases; on the other hand, it detects 62% of the strong connection, but in this case it produces 66% of false positives, which is comprehensible given the small number of these connections.

More tests were developed on the intervals: 1–10 links (about 9% of the connections), 11–50 (2%), 51–250 (0.6%), above 250 links (0.10%). Even if some of them represent very small shares of dyads, the model is able to correctly detect a significant number of the cases. For instance, the model is able to find 36% of the connections between 51 and 250 links, while producing seven false positive every ten predictions.

Table 4
Web link distribution along antecedents.

	Mean links		Total	Ranges in the number of links of the dyads							Share of active dyads	Strong/weak ratio
				0	1–10	11–50	51–250	251–1000	1001–10,000	>10,000		
Country												
Same country	19.1	No. dyads	144,556	98,794	29,106	9955	5196	1166	315	24	32%	0.57
		Links	2,766,230	–	96,827	235,335	563,345	524,743	840,080	505,900		
Different countries	1.0	No. dyads	1,249,024	1,127,659	98,041	18,955	2854	279	51	4	10%	0.23
		Links	1,303,894		304,569	387,757	276,449	122,869	136,550	75,700		
Distance between sending and target HEIs												
<100 km	44.5	No. dyads	17,679	11,869	2243	1489	799	306	116	9	33%	1.21
		Links	787,029		7579	34,098	92,306	146,065	287,210	212,500		
100–500 km	8.9	No. dyads	164,491	126,545	26,037	7388	3537	655	152	9	23%	0.45
		Links	1,472,181	–	86,182	171,847	379,747	288,134	410,650	143,600		
500 km	1.5	No. dyads	1,211,410	1,088,039	98,867	20,033	3714	484	98	10	10%	0.25
		Links	1,810,914	–	307,635	417,147	367,741	213,413	278,770	225,500		
PhD												
Both PhD awarding	7.0	No. dyads	533,630	406,885	91,970	25,531	7535	1,350	332	27	24%	0.38
		Links	3,709,563		307,821	559,239	782,625	604,458	884,520	570,900		
Sender and/or target with no PhD	0.4	No. dyads	859,950	819,568	35,177	3379	515	95	34	1	5%	0.11
		Links	360,561		93,575	63,853	57,169	43,154	92,110	10,700		
Subject similarity												
High (0.50–1)	6.4	No. dyads	447,617	340,370	78,371	21,132	6343	1129	255	17	24%	0.37
		Links	2,851,679	–	259,355	461,478	666,175	503,751	655,920	305,000		
Low (0–0.50)	1.3	No. dyads	945,963	887,264	48,776	7778	1707	316	111	11	6%	0.20
		Links	1,218,445	–	142,041	161,614	173,619	143,861	320,710	276,600		
Receiver size												
Size > 320	5.5	No. dyads	696,790	558,260	102,318	26,770	7679	1389	347	27	20%	0.35
		Links	3,828,437		334,654	581,875	797,101	620,097	923,810	570,900		
Size < 320 (median)	0.3	No. dyads	696,790	668,193	24,829	2140	371	56	19	1	4%	0.10
		Links	241,687		66,742	41,217	42,693	27,515	52,820	10,700		
NUTS region												
Same NUTS region	69.9	No. dyads	10,010	5268	2661	1,076	643	241	113	8	47%	0.78
		Links	699,629	–	9171	24,855	73,841	119,142	283,620	189,000		
Different NUTS region	2.4	No. dyads	1,383,570	1,222,366	124,486	27,834	7407	1,204	253	20	12%	0.29
		Links	3,370,495	–	392,225	598,237	765,953	528,470	693,010	392,600		
Legal status												
Same legal status	4.3	No. dyads	908,604	760,684	111,242	27,098	7818	1,380	356	26	16%	0.33
		Links	3,863,073		357,104	587,343	815,168	617,078	946,480	539,900		
Different legal status	0.4	No. dyads	484,976	466,950	15,905	1812	232	65	10	2	4%	0.13
		Links	207,051	–	44,292	35,749	24,626	30,534	30,150	41,700		

Table 5
Negative binomial regression models: active connections in the section above, zero values below.

	NULL model (1)			Country model (2)			Complete model (3)		
	Estimate	Std. error	Pr	Estimate	Std. error	Pr	Estimate	Std. error	Pr
(Intercept)	−13.73	33.6		−16.29	91.07		−15.52	43.93	***
Country				2.17	0.014	***	1.20	0.016430	***
Receiver size							0.0005	0.000003	***
Sender size							0.0005	0.000003	***
Distance							−0.0005	0.000006	***
Subject							1.07	0.022930	***
PhD							0.44	0.016520	***
NUTS regions							1.94	0.042540	***
Log (theta)	−18.50	33.6		−19.99	91.07		−17.25	43.93	
Zero hurdle model coefficients (binomial with logit link)									
(Intercept)	−2.00	0.00262	***	−2.24	0.00303	***	−5.36	0.013550	***
Country				1.47	0.00642	***	1.74	0.010500	***
Receiver size							0.0007	0.000004	***
Sender size							0.0007	0.000004	***
Distance							−0.0002	0.000004	***
Subject							2.83	0.013110	***
PhD							1.38	0.007184	***
NUTS regions							1.20	0.025920	***
Theta: count		0			0			0.00E+00	
Number of iterations		20			20			50	
Log-likelihood:		−1.047E+06 on 3 Df			−1.009E+06 on 5 Df			−8.34E+05 on 17 Df	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘’ 1.

Table 6
Negative binomial regression model: measure of statistical performance.

	Non-active: “zero”	Active	’Strong’: above 10 links	1–10 links	11–50 links	51–250 links	Above 250 links
Sensitivity	85%	75%	62%	49%	43%	36%	15%
Specificity	75%	85%	97%	86%	97%	99%	100%
Positive predictive value	96%	41%	34%	26%	21%	29%	35%
Negative predictive value	41%	96%	99%	94%	99%	100%	100%

Table 7
Negative binomial regression model: comparing the impact of the variables.

Variable	Delta	Expected value change	
		Proportion in *N* of links	Likelihood of linking
Country	1–0	3.3	5.7
Subject	1–0	2.9	17.0
PhD	1–0	1.6	4.0
NUTS region	1–0	6.9	3.3
Distance	+1000 km	−0.6	−0.8
Receiver size	+1000 units of staff	1.7	2.1
Sender size	+1000 units of staff	1.6	2.1

The model coefficients are informative of the strength of the impact of antecedents. For instance, if the linked HEI “B” is in the same country of the linking HEI “A”, it is predicted that A will send $e^{+1.20} = 3.3$ times more links to B than to a HEI “C” in a different country, while A is $e^{+1.74} = 5.7$ times more likely to link B than a HEI “C” in a different country (zero hurdle model).

First, the impact on the number of links is analyzed. Among the variables ranging from 0 to 1, the NUTS region is the strongest (Table 7), followed by the country and by the subject similarity – a HEI is expected to send to another HEI with the same discipline profile almost three times the links sent to a HEI with a completely different profile – and by the PhD. The NUTS variable is also very strong when compared to size and distance: being in a different region produces a reduction in the expected number of links corresponding to having 3.812 less units of academic staff or being 3.756 km further. This confirms previous results on the importance of national and regional structures, i.e. of socio-institutional spaces, in interlinking patterns, while only the very large generalist doctorate awarding universities tend to display significant interlinking levels across countries. It is important to note that country and, especially, NUTS have a strongly focused impact, as they occur in 10% and 0.7% of the cases, whereas the PhD variable is equal to 1 for 38% of the dyads.

When analyzing the impact on likelihood of linking results are similar. All the coefficients, with the exception of the distance’s and the NUTS region, become stronger. This hints to the fact that belonging to the same institutional space

Table 8
Negative binomial regression models on the intensity of connections among universities in the Leiden ranking.

	NULL model (1)			Model (2)			Complete model (3)		
	Estimate	Std. error	Pr	Estimate	Std. error	Pr	Estimate	Std. error	Pr
(Intercept)	4.05908	0.01	***	1.88	0.04	***	−1.41	0.069	***
Country				1.90	0.03	***	2.05	0.026	***
Receiver size				0.00039	0.000008	***	0.0004	0.000007	***
Sender size				0.00030	0.000008	***	0.0003	0.000007	***
Distance				−0.00054	0.000012	***	−0.0004	0.000012	***
Subject				0.62	0.03	***	0.71	0.033	***
Receiver quality (IF)							1.02	0.036	***
Sender quality (IF)							1.80	0.036	***
Dispersion parameter	0.2854			0.4544			0.4924		
AIC	310,033			289,970			286,761		
Theta	0.28545			0.45439			0.49237		
St. Error	0.00189			0.00323			0.00355		
2 × log-likelihood:	−310,028.912			−289,956.15			−286,742.653		

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1.

Table 9
Negative binomial regression model: comparing the impact of the variables on the connections between universities in the Leiden ranking.

	Non-active: “zero”	Active	‘Strong’: above 10 links	1–10 links	11–50 links	51–250 links	Above 250 links
Sensitivity	0%	100%	92%	31%	67%	42%	51%
Specificity	100%	0%	36%	82%	48%	89%	96%
Positive predictive value	0%	86%	54%	55%	36%	34%	27%
Negative predictive value	86%	0%	85%	63%	77%	92%	98%

(country, region or HEI type) largely determines the probability of interlinking, while factors like distance and size are more important in moderating the strength of the connections.

The high value of the subject coefficient can be explained by the fact that a consistent number of dyads show a very small discipline similarity (14% of the sample is equal to “0”), and these dyads have a very low probability of interlinking even when they are in the same country or geographically near. For instance, the share of active dyads among HEIs in the same country is 32% but it drops to 7% if they have a discipline similarity equal to “0”. If two HEIs have a completely different subject profile, the probability of interlinking will thus be very low independently of other factors.

5.1. Studying the intensity of connections among high ranked universities

In order to analyze for the impact of productivity and quality measures, we test the model on the sample of HEIs in the Leiden ranking; HEIs in this sample are larger than the general sample (2159 academic units) and all PhD awarding. Also the intensity of connections is very different, as the non-active connections are a minority (14%), whereas active connections are the large majority: 1–10 links (41%), 11–50 (30%), 51–250 (12%) and above 250 links (3%).

In this case a negative binomial regression model has been used, as the number of zeros is low. The model including the weighted impact factor is significantly superior to the one without quality (Table 8). Other tests have been run with other indicators of research performance, such as of productivity, but they perform worse.

As shown by Table 9, the model is not able to detect non active connections as their share is quite small in the sample, while its performance is much better as to the capability to detect connections above 10 links (45% of the sample), where it can identify 36% of the cases (specificity) and correctly predict 54% of the positives cases. Expectedly, the level of research quality becomes relevant as a predictor, especially when the number of links is rather high, as displayed by the good performance in identifying the dyads with more than 250 links.

Table 10
Negative binomial regression model: comparing the impact of the variables.

Variable	Delta	Number of links
Country	1–0	7.7
Subject	1–0	2.0
Distance	1000 km	0.64
Receiver size	1000 units of staff	1.7
Sender size	1000 units of staff	1.3
Receiver reputation	1 IF	2.8
Sender reputation	1 IF	6.0

As shown by Table 10, two universities in the same country are expected to exchange almost eight times more links that if they are located in different countries. The importance of subject similarity is much weaker (around two). The country variable is also strong when compared to the size and the distance.

Research quality appears very important as well. The *quality* factor has a range of 1.18, from 0.51 to 1.69. Thus, when the differences in *quality* are large, the overall impact is even stronger than being in different countries. Interestingly the quality of the sender has a larger impact than the quality of the receiver. This draws to very strong interlinking patterns among the highly reputed HEIs, as the expected number of links between two HEIs with IF 1.5 is more than 15 times the one between two HEIs with IF 0.5.

These results may be summarized as follows. First, research productivity is not a relevant factor impacting number of weblinks because most of them are not directly related to research production. On the contrary, research quality has a very strong impact through the reputational structure of HE systems, which are characterized by a core composed by the most reputed HEIs strongly interconnected among them. Accordingly, the individual level of reputation of sender and receiver is more important than their difference.

In turn, these results lend to the hypothesis that HEIs of lower reputation in the periphery are preferentially connected with the center through highly reputed bridging institutions, while being less connected among them. The fact that this structure emerges in a European sample, despite national differences, suggests that most European higher education systems display this core-periphery structure and that in most of them reputation is a determining factor of centrality.

6. Discussion

Before discussing the implications of our results, it is important to acknowledge the limitations of the study. The literature on web links among HEI supports the claim that they reflect underlying inter-organizational relations, but involve a variety of motivations. Accordingly, one needs to be careful in interpreting tie strengths, as different numbers of web links might be generated by different types of relationships. Our data provide a single time window; moreover, one has to take in mind that data have been extracted with a specific search engine, whereas alternative search engines may have a different country coverage (Barjak & Thelwall, 2008). Some studies show however that the relative strength of connections have rather stabilized, thus increasing the longer-term validity of webometrics (Payne & Thelwall, 2007).

We also consider that some macro-trends, like the increasing diffusion and content of researchers' webpages, might change in the future the characteristics of the web presence of HEIs, implying that the meaning of connections become more aligned with researchers' activity and less with institutional ones (Barjak, Li, & Thelwall, 2007). Only longitudinal studies allow addressing these concerns. Finally, we did not consider France, as it was not covered by the EUMIDA dataset and this might affect some results given the size of the country – EUMIDA identified about 120 research-active French HEIs, i.e. about 10% of the sample considered.

Despite these limitations, our study goes beyond prior work both methodologically as well as substantively. At the methodological level, we showed that binomial regression provides a suitable tool for analyzing weblinks which is both conceptually and statistically more suited to their characteristics than ordinary regressions and which is less sensitive to the presence of outliers. Moreover, these allow a more accurate interpretation of results in terms of predictive ability of weblink counts and of the strength of the coefficients.

This result was also due to a second novelty, namely matching the weblink data with structural data of a large sample of European HEIs; as web links cannot be directly related to a specific kind of relationships, their interest is limited if these are not connected to some other information, like structural data and/or other types of relational data. If compared to previous national-level studies, the size and composition of the sample allows for more statistical robustness and generality of the results, which are not tied to specific national structures (i.e. a concentration of the large and most reputed HEIs in some regions).

At the substantive level, our study confirms previous work showing that counts of weblinks can be predicted with reasonable precision from what social network theory indicates as antecedents of HEI relationships, thus confirming insights from motivation studies that weblinks are not technical artifacts. This insight should foster a broader use of weblink data to analyze relational structures of higher education systems.

In this respect, we provided three advances: first, our analysis extends well beyond research intensive universities to include most of European higher education institutions, thus allowing to observe interlinking patterns also in the periphery of the system. While top-ranked universities are interconnected across Europe (84% of the dyads in the Leiden ranking being active), the broader HEI system is more sparse and fragmented according to national (and regional) spaces; this confirms insights that there are different levels of HE networks, with a European-level network of research universities connected through a hub and spoke structure to national networks of less reputed HEIs.

Second, we were able for the first time to systematically compare the strength of the antecedents, showing that three main factors structure HE relational spaces, namely the country and the region, the subject domain and the divide between PhD-awarding and non-PhD awarding institutions. Two further factors act as moderators of the strength of interlinking, namely the HEI size and their geographical distance. Third, the test on the Leiden sample demonstrates that, while institutional spaces are most important in the broader HEI system, among highly reputed universities research reputation is the strongest factor influencing interlinking patterns; this empirically confirms insights concerning the core-periphery structure of HEI networks and the fact that research reputation is the discriminating factor to belong to the core.

Finally, we hint to four directions for future work. First, it would be important to model the exponential part of the weblink distribution (i.e., the 2000 dyads above 250 links), as there are indications that many of the high counts are not due to technical artifacts, but to substantive factors; as these cases are relatively few, qualitative investigations might hint to relevant factors which allow to better estimate of the probability of getting high values. Second, while our focus was on general patterns, we suggest investigating also national differences in the importance of antecedents, i.e. including country effects or additional variables characterizing groups of countries and measuring goodness of fit at national level; this might provide useful insights on the influence of national higher educational policies on the structure of HEI networks. Third, the hierarchy of networks emerging from our data, as well as the core-periphery structure of national higher education systems, deserves careful investigation. Accordingly, testing the predictive ability of antecedents on measures of network positions like centrality and connectivity could allow to better understand the factors determining HEI centrality in different levels of networks. Fourth, and in conclusion, the results we have presented stand on an analytically convenient hypothesis of independence between dyads. Future research should establish the extent to which this hypothesis is also empirically plausible.

Acknowledgments

We thank Francesca Pallotti for comments on preliminary versions of the article. We thank Marco Calderisi for his help with the statistical tests and Michele Seeber for the support to the Python scripts. A preliminary version of this paper was presented to the ENID/STI conference, Rome, September 2011. We also thank two anonymous referees for their useful comments.

References

Aguillo, I., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetric indicators. *Journal of the American Society for Information Science and Technology*, *57*(10), 1296–1302.

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

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country – The case of Israel. *Scientometrics*, *59*(3), 391–403.

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.

Barjak, F., Li, X., & Thelwall, M. (2007). Which factors explain the Web impact of scientists' personal homepages? *Journal of the American Society for Information Science and Technology*, *58*(2), 200–211.

Barjak, F., & Thelwall, M. (2008). A statistical analysis of the web presences of European life sciences research teams. *Journal of the American Society for Information Science and Technology*, *59*(4), 628–643.

Bonaccorsi, A., Lepori, Brandt, De Filippo, Niederl, Schmoch, et al. (2010). Mapping the European higher education landscape. In *New insights from the EUMIDA project*.

Brass, D., Galaskiewicz, J., Greve, H. R., & Tsai, W. (2004). Taking stock of networks and organizations. A multilevel perspective. *Academy of Management Journal*, *47*(6), 795–817.

Braun, T., & Glänzel, W. (1996). International collaboration: Will it be keeping alive East European research? *Scientometrics*, *26*(2), 147.

Breschi, S., & Lissoni, F. (2004). Knowledge networks from patent data. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 613–644). Dordrecht: Kluwer.

Broder, A., Kumar, Maghoul, Raghavan, Rajagopalan, Stata, et al. (2000). Graph structure in the Web. *Computer Networks and ISDN Systems*, *30*, 209–320.

Burris, V. (2004). The academic caste system: Prestige hierarchies in PhD exchange networks. *American Sociological Review*, *69*(2), 239–264.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge: Cambridge University Press.

Clark, B. R. (1983). *The higher education system. Academic organization in cross-national perspective*. Berkeley: University of California Press.

de Boer, H., Enders, J., & Leisyte, L. (2007). Public sector reform in Dutch higher education: The organizational transformation of the university. *Public Administration*, *85*(1), 27–46.

Deiaco, E., Holmén, M., & McKelvey, M. (2010). What does it mean conceptually that universities compete? In M. McKelvey, & M. Holmén (Eds.), *Learning to compete in European Universities* (pp. 300–328). Cheltenam: Edward Elgar.

Doz, Y. L., Oik, P. M., & Ring, P. S. (2000). Formation processes of R&D consortia: Which path to take? Where does it lead? *Strategic Management Journal*, *21*, 239–266.

Evans, T. S., Lambiotte, R., & Panzarasa, P. (2011). Community structure and patterns of scientific collaboration in business and management. *Scientometrics*, *89*(1), 381–396.

Glänzel, W. (2001). National characteristics in international scientific co-authorship. *Scientometrics*, *51*(1), 69–115.

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.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *The American Journal of Sociology*, *91*(3), 481–510.

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science*, *30*(5), 436–447.

Heimeriks, G., & Van den Besselaar, P. (2006). Analyzing hyperlink networks: The meaning of hyperlink based indicators of knowledge. *Cybermetrics*, *10*(1).

Heller-Schuh, B., Barber, M., Henriques, L., Paier, M., Pontikakis, D., Scherngell, T., et al. (2011). *Analysis of networks in European framework programmes (1984–2006)*. Luxembourg: Publications Office of the European Union.

Jones, B. F., Wuchty, S., & Uzzi, B. (2008). Multi-university research teams: Shifting impact geography, and stratification in science. *Science*, *322*(5905), 1259–1262.

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.

Merton, R. K. (1968). The Matthew effect in science. The reward and communication systems of science are considered. *Science*, *159*(3810), 56–63.

Mullahy, J. (1986). Speciation and testing of some modified count data models. *Journal of Econometrics*, *33*, 341–365.

Opsahl, T., Colizza, V., Panzarasa, P., & Ramasco, J. J. (2008). Prominence and control: The weighted rich-club effect. *Physical Review Letters*, *101*, 168702.

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.

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

Powell, W., Koput, K., & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, *41*(1), 116–145.

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.

Schubert, A., & Braun, D. (1990). International collaboration in the sciences 1981–1985. *Scientometrics*, *19*(1–2), 3–10.

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., & Harries, G. (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), 149–159.

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

Thelwall, M., & Zuccala, A. (2008). A university-centred European Union link analysis. *Scientometrics*, *75*(3), 407–420.

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. (2008). *The new invisible college*. Washington, DC: Brookings Press.

Weick, K. (1976). Educational organizations as loosely coupled systems. *Administrative Science Quarterly*, *21*(1), 1–19.

White, H. C. (2002). *Markets from networks. Socioeconomic models of production*. Princeton, NJ, USA: Princeton University Press.

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.

Wuchty, S., Jones, B., & Uzzi, F. B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039.