UNIVERSITY OF GRANADA
Faculty of Economics and Business Studies
Department of Accounting and Finance

PhD Thesis
abridged version in English

APPLICATION OF WEBOMETRIC TECHNIQUES TO THE STUDY OF ACCOUNTING AND FINANCIAL VARIABLES

Esteban Romero Frías

Supervisors:
Liwen Vaughan // Lázaro Rodríguez Ariza

GRANADA, 2010

Table of Contents

1 INTRODUCTION

This is an abridged version in English of the thesis presented by Esteban Romero Frías in the Department of Accounting and Finance of the University of Granada in 2010.

This thesis constitutes an overview of the Webometric methodology to perform quantitative research on business Web sites. So far, most of the research on commercial Web sites using Webometric methods have been developed from an information science perspective. However, there is a promising field to apply these techniques to different types of business issues.

Apart from this brief introduction, this document contains three main sections:

- Section 2, which addresses theoretical issues;
- Section 3, which includes several empirical studies; and
- Section 4, which develops the general discussion, limitations and future research of the thesis.

2 THEORETICAL APPROACH

This section includes a paper that summarizes the webometric approach to business studies, based on link impact analysis and co-link analysis. Both techniques are used in the empirical studies included in chapter 3.

The reference of the following paper is:

> ROMERO-FRÍAS, E. (2009) "Googling Companies – A Webometric Approach to Business Studies". *Electronic Journal of Business Research Methods*, 7(1): 93-106, available at: http://www.ejbrm.com/vol7/v7-i1/v7-i1-art10-abstract.htm

Googling Companies -

A Webometric Approach to Business Studies

Abstract: So far Internet studies have focused mainly on using website content for gathering business information, however web hyperlinks have not been exploited enough for business purposes yet. Webometric techniques are based on the exploitation of information contained in the hyperlinks that connect the different documents contained on the Web. Webometrics could be considered as a new discipline that applies bibliometric techniques to the quantitative study of the Web, but also a discipline that progressively develops its own concepts and methodology. So far studies in this field have focused on academic and scholarly web spaces; however this methodology is equally applicable to commercial sites which are more predominant on the Web.

This paper is intended to show how webometric techniques could be applied to business and management studies. Therefore, it describes a number of basic concepts and techniques and the way in which they have been applied to these fields so far. Firstly, some studies found that the number of links pointing to companies' websites correlates significantly with the business performance measures of the entity. This finding suggests that links to a website could be used as a timely indicator of business performance. Secondly, the examination of co-links, which refers to webpages that links two business sites simultaneously, have been used for competitive intelligence purposes. These studies are based on the idea that the number of co-links to the websites of a pair of companies is a measure of the similarity between them. For instance, this similarity measure between companies in the same industry can provide information about their competitive positions. Finally, motivations for the creation of hyperlinks to business sites could be analysed through a content analysis approach in order to get confirmation about the business relevance and nature of links. This view complements the quantitative perspective to link and co-link research, providing a brand new approach to business studies.

Keywords: Web mining, webometrics, business intelligence, business management, internet studies

1. Researching the Web: Why hyperlinks do matter

Since its creation in 1989, the World Wide Web (the Web) has revolutionised the Internet, facilitating the access to information to many potential users. Two decades later, the Web has become part of the daily lives of many people all over the world, causing deep social transformations that social scientists struggle to understand. Moreover, for the past five years, the Web has undergone significant changes by the popularisation of the so-called Web 2.0 (O'Reilly, 2005). This has provoked a democratisation of the information creation tools in such a way that millions of people have started to participate in a global conversation based on the use of no cost, user friendly, multimedia and Web-based software (Jenkins, 2006). The blossoming of publishing on the Web has made the media more difficult to understand, always in a process of constant change, described by many as chaotic, uncontrolled and of poor quality (Keen, 2007). However, the Web and the Internet as a whole are the largest repositories of information ever known in history.

In 2008, the official Google blog (2008) reported that the number of pages indexed had grown to more than one trillion (as in 1,000,000,000,000) unique URLs (Uniform Resource Locators). Technorati (2009), the main blog search engine, reported to be tracking on the order of 133 million blogs by the end of 2008. Most, if not all, of the human activities, whether they are political, social, educational or economical are reflected on the Internet, and some have developed native online phenomena that would not exist in the offline world.

The available collection of information makes it possible to define the Web as an enormous unstructured and heterogeneous database that, despite its appearance, is not randomly built. This implies that this could be exploited from different perspectives (based on content, structure and user behaviour) in order to study unique online phenomena or offline phenomena reflected in the Web. A basic component of the array of information available is the hyperlink.

A hyperlink is a reference or navigational element in a document to another section of the same document or to another document that may be on or part of a different domain. Hyperlinks represent the hidden structure of the Web connecting different sites and webpages that would stay isolated unless the specific URL is known (Berners-Lee, 1999). They can be regarded as an endorsement of a target page, especially if the creator has placed that link because it points to a useful or relevant resource. The creation and exploitation of hyperlinks are not an irrelevant phenomena, but imply significant social repercussions (Turow and Lokman, 2008). It is not meaningless that currently most of the search engines use link analysis as part of their algorithms to crawl websites and to rank the pages (Batelle, 2005).

Google’s market dominance derives originally from the technological lead established by the Pagerank system introduced in 1998 (Brin and Page, 1998). This Pagerank is based on hundreds

of factors that have changed over time in order to avoid manipulation, but basically follows two principles:

- Webpages that have more links pointing to them are considered to be more relevant.
- Not all the hyperlinks have the same value. Those links established by relevant webpages are more valuable than others from less well known webpages.

However, hyperlinks are not a perfect source of evidence because many of them have not been created by a thoughtful process in order to endorse or discredit a webpage. Many links are created for navigational purposes within a site; others are just automatically created by content management systems or, in the worse case, they are just spam or lists of URLs created to perform better in the search engines rankings.

Hyperlinks constitute the raw basic material of quantitative research in the Web, as performed by Webometrics.

2. An introduction to Webometrics

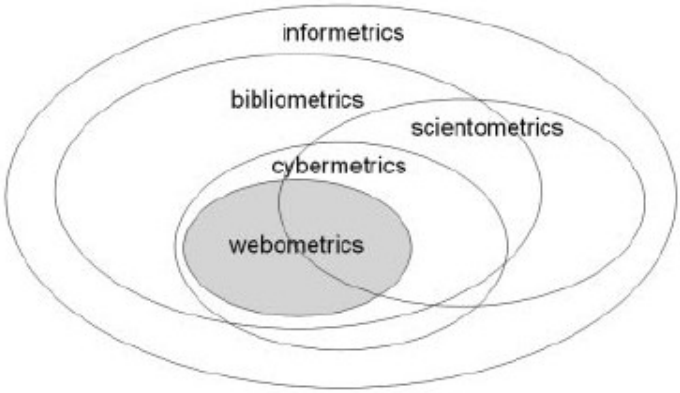
The idea of the Web as a distributed database that is exponentially growing over time has been appealing for Data Mining research. There is a great potential for analysis by mining the website content, document structure, site relations, and user behaviour (Benoît, 2002). In this context, Webometrics has developed into a new discipline that studies Web-phenomena from a quantitative point of view.

2.1. Definition

The origin of Webometrics can be found in the field of Information Science. Thelwall, Vaughan and Björneborn (2005: 81) point out that the discipline “emerged from the realization that methods originally designed for bibliometric analysis of scientific journal article citation patterns could be applied to the Web, with commercial search engines providing the raw data”. In fact, the idea that a link pointing to a webpage means a 'vote' to that webpage or document is based on bibliometric methods to rank scientific production (Garfield, 1979). The term Webometrics was first coined by Tomas Almind and Peter Ingwersen in 1997 and seems to be widely accepted by the research community together with the term Cybermetrics. Björneborn (2004) defined both terms by limiting their research areas. Webometrics is "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches" (in Björneborn and Ingwersen, 2004: 1217), while Cybermetrics does the same but on the whole Internet. Hence, Cybermetrics is more focused on the study of non web-

based Internet phenomena, e.g. emails, chat, newsgroup studies, etc. Figure 1 shows the location and overlapping of these disciplines in the general context of Information Science.

Figure 1: Webometrics and Cybermetrics in the context of Information Science (Björneborn and Ingwersen 2004: 1217)



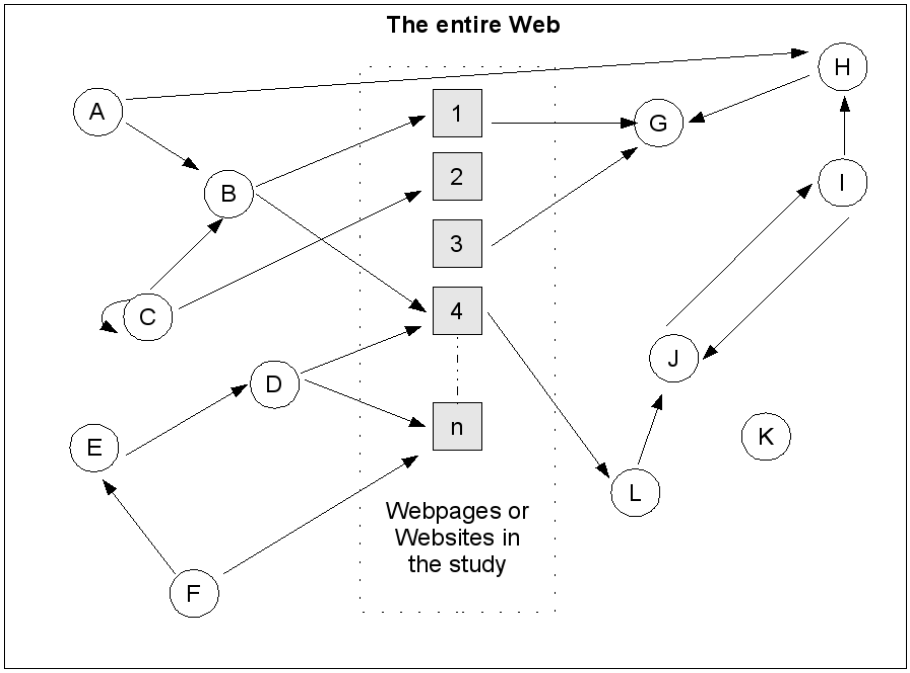
Recent developments within the field suggest a move in the scope of the definition into a more general social science research approach instead of an approach that is mainly based on a informetric and bibliometric perspective. Thelwall (2009: 6) defines Webometrics as “the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study”. Interdisciplinary research is getting more significant by enlarging the types of subjects of study and the techniques used. This evolution aligns with the definition of Internet research given by Hine (2008: 537): “Internet research itself is not a discipline but an interdiscipline, a field or a research network populated by heterogeneous perspectives.”

2.2. Basic concepts

Björneborn and Ingwersen (2004) carried out the first attempt to develop a consistent terminology on the webometric field. Some years later, Thelwall and Wilkinson (2008) proposed a generic lexical framework that, based on the previous work, intended to unify and extend existing methods through abstract notions of link lists and URL lists.

Figure 2 shows a diagram of the Web where the circles represent different types of nodes (websites, webpages, etc.) and the arrows connections between them. The squares within the dashed line rectangle are nodes that are considered for analysis in a given research. They are specially useful to illustrate a specific type of analysis, the so called, co-link analysis.

Figure 2: Diagram with different types of links existing in the Web.



Letters in the diagram represent any type of document in the Web, whether it is a webpage or a website, for instance. The following basic webometric terms (Björneborn and Ingwersen, 2004) can be explained by reference to Figure 2:

- **Inlink**: B has an inlink from A.
- **Outlink**: A has an outlink to B.
- **Self-link:** C has a self-link.
- **Page or site isolated**: K is isolated as it does not have any inlinks or outlinks.
- **Reciprocal links**: I and J have reciprocal links.
- **Transversal link**: A has a transversal outlink to H. This type refers to a link that joins to different areas of the Web that are not well interconnected.
- **Co-inlinks**: 1 and 4 have a co-inlink, as B links both of them simultaneously.
- **Co-outlinks**: G has a co-outlink, as 1 and 3 are linking to it.

This paper basically focused on inlink and co-inlink analysis, as most of the webometric research on business is based on these two concepts. Co-inlink analysis is referred simply as co-link analysis later in this paper.

2.3. Methodology

Before reviewing the application of webometric techniques to business research, this paper offers a general overview of the methodology. The application of bibliometric techniques to the Web had derived from the similarities between hyperlinks and academic citations, considering that both point from a source document to a target document. Nevertheless, important differences exist.

In this paper, we briefly describe three main types of techniques based in the use of commercial search engines to gather raw data. For a more systematic approach, a complete and didactic reference is Thelwall (2004; 2009).

2.3.1. Web impact analysis

Web impact analysis provides evidence for the impact or the spread of ideas, brands, organisations, etc. on the Web by measuring and analysing the URLs retrieved by commercial search engines in response to a specific query. This approach is especially useful as an exploratory approach for further research, although it has some significant drawbacks. For instance, one such problem is the extent to which the keyword used matches or does not the subject under research. This technique is the only one in this paper that is not based on the analysis of hyperlinks, however as it is the most intuitive one is appropriate to introduce the subject.

The main problems come from keywords that are too common or have different meanings, resulting in a wide variety of results that do not necessarily match the issue under study. In this case, it is necessary to filter out false matches. However, this is not always possible due to search engines limititations (see section 2.4).

Practical example: To carry out a basic impact analysis of two commercial banks in the UK, queries [“Royal Bank of Scotland”] and [Barclays] could be searched on Google, Yahoo! or Bing (MSN). Nevertheless, the number of matches is so high that the researcher might need to refine the search in order to focus on a specific issue concerning the companies. For instance, to explore the effects of the financial crisis on the banks, queries could be formulated such as [“Royal Bank of Scotland” AND “financial crisis”] and [Barclays AND “financial crisis”].

2.3.2. Link impact analysis

Link impact analysis is based on the comparison of the number of webpages or websites that link to a set of webpages or websites under research. The purpose of this type of research is,

according to Thelwall (2009: 28), “to evaluate whether a given website has a high link-based web impact compared to its peers”. Also inlink counts can be an indirect gauge of other attributes of the organization represented by the website. For instance, this has been traditionally used, within the academic field, as a potential estimator of research performance (Smith and Thelwall, 2002). This method is more accurate than the previous one because false matches are surely avoided. However, other problems need to be taken into account, e.g. the search engines set restrictions on search of inlinks (see section 2.4), or the company is using more that one corporate URL.

Practical example: Following up the aforementioned example, to carry out a link impact evaluation of the two companies, the queries that should be used on Yahoo! (the search engine with the best array of functions for this purpose) would be as follows: [linkdomain:rbs.com -site:rbs.com] and [linkdomain:barclays.com -site:barclays.com]. The results would be the estimated total number of links that point to the specific domain except the links that come from the same domain or self-links.

2.3.3. Co-link analysis

Co-link analysis could be classified as a type of link relationship mapping techniques (Thelwall, 2009). These are based on the link data that interconnect a set of websites in different ways in order to draw a diagram that illustrates the relationships between them. In particular, co-link analysis is based upon the number of webpages that link at the same time two webpages or sites belonging to the group of entities under study. This description fits the concept of co-inlink previously explained. As we mentioned before, co-link is often used as equivalent to the co-inlink concept. It is the case in this paper.

Co-links are analogous to the bibliometric concept of co-citation (Small, 1973). Co-link analysis has also been demonstrated to be a useful tool to reveal the cognitive or intellectual structure of a particular field of study (Zuccala, 2006). This method is particularly useful when websites interlink each other very rarely. It is the case of commercial websites that scarcely link the website of a competing company, especially when they are in the same industry (Vaughan, Gao and Kipp, 2006). The explanation to this could be that companies seem to avoid diverting web traffic to competitors (Shaw, 2001). Moreover, as Vaughan (2006) points out, co-link data are more robust than inlink data as the former are less easily manipulated. Multidimensional scaling and network diagrams are often used to show the data gathered and to interpret results.

Practical example: The query [linkdomain:rbs.com -site:rbs.com linkdomain:barclays.com -site:barclays.com] would retrieve the estimated total number of links that point simultaneously to both domains except the links that come from the same domain or self-links.

To conclude this section, some final considerations need to be done. Due to the origins of the discipline, so far the majority of webometric research has been carried out in the academic field. Thelwall, Vaughan and Björneborn (2005: 113) acknowledge this situation and point out that “This is ironic given that the Web is dominated by commercial sites”. Webometrics is progressively enlarging its scope, focusing upon political websites (Foot and Schneider, 2006; Park and Thelwall, 2008), social networking (Thelwall, 2008b; 2008c; 2008e) and commercial websites (see section 3).

2.4. Collecting data using Commercial Search Engines

Search engines are crucial for webometric research, because their databases are the source of information that cover most of the Web. Despite the fact that personal web crawlers can be used to automatically download pages and extract their links, commercial search engines have been used extensively for research especially when large areas or potentially the whole Web are the object of the study.

In order to perform a better research using commercial search engine data, it is fundamental to get a good understanding of the industry context, the advanced functions offered and the limitations.

A feature of the search-engine market is the oligopoly of three search-engine operators Google, Yahoo! and Live Search (Microsoft) which, from a global perspective, share the majority of the generalistic search-engine market. In specific areas of the Web the are other players such as Technorati for searching blogs. Search engine industry is under a constant process of change and innovation. This issue has been treated by many papers in recent years (Bar-Ilan, 2004; Vaughan and Thelwall, 2004; Lewandowski, Wahlig and Meyer-Bautor, 2006; Evans, 2007; Thelwall, 2008d).

As already mentioned, commercial search engines are the only source of data that covers the entire Web. However there are some significant limitations derived from the use of commercial search engines:

- Search engines do not index the entire Web (Sherman and Price, 2001; Bar-Ilan, 2004; Thelwall, Vaughan and Björneborn, 2005).
- Ranking systems eliminate similar or identical pages in their results, in order to avoid providing useless information (Gomes and Smith, 2003; Thelwall, 2008a).
- Crawling and reporting algorithms are commercial secrets and, therefore the exact criteria used to rank the information is unknown (Thelwall, Vaughan and Björneborn, 2005).
- The total number of results offered by search engines are estimates as they use algorithms

that prioritise response time rather than exhaustiveness (Björneborn and Ingwersen, 2001).

- Results can be subject to national or language biases (Vaughan and Thelwall, 2004).
- The results can fluctuate and change over the time. In addition, only a few number of pages are accessible (usually just a maximum of 1000).

Commercial search engines are the best and unique source of information we have for certain types of webometric research, however they are not designed with this academic purpose and the results are not as exhaustive as we would desire. At this moment, Yahoo! is the search engine that is more useful for webometric research (Thelwall, 2008d). Yahoo! inlink data can be gathered in two different ways, through the general Yahoo! search engine and through Yahoo! Site Explorer. Despite the latter specializing in web structure information, complex queries can only be submited through the general Yahoo! search engine interface.

Nevertheless, collecting data can be a very time consuming process if using the web interface. This problem could be overcome by specialized software based on the application programming interfaces (API) developed by search engines and other services on the Web.

3. A webometric approach to business studies

Internet research applied to companies can provide new insights on Competitive Intelligence (CI) in order to help companies generate and maintain a competitive advantage (Porter, 1980). CI consists of a systematic plan to obtain and analyse information about competitors and general trends in the industry (Kahaner, 1996). The abundance of information available in the Internet generates new opportunities and challenges for businesses that need to monitor changes around them in order to compete in better conditions.

In the last decade, a few Web hyperlink analysis have been conducted for CI purposes with promising results (Reid, 2003; Tan *et al*., 2002). For example, Reid (2003) proposed a link analysis method which analysed a particular website, but not the global context of the Web. Over the last 5 years, Vaughan and colleagues have investigated the relationships between inlink counts and business performance measures as well as the application of co-link research to companies' websites.

3.1. Do inlinks and financial variables correlate?

Vaughan and colleagues' research on commercial websites has explored quantitative relationships between inlink counts and business performance variables. Vaughan and Wu (2004) made the first attempt to prove the hypothesis that the number of inlinks to commercial sites correlates with

financial variables. This study tests the hypothesis in two groups of Chinese companies. Group 1 is made up of China's top 100 Information Technology (IT) companies and group 2 is comprised of top 100 privately owned companies.

The IT industry was selected because companies in this industry, by the nature of their activity, are likely to be leaders in utilizing the Web for commercial purposes. Experience has shown that different industries, have distinct patterns in the use of web technologies and therefore only homogeneous companies are likely to be comparable. Group 1 constitutes a homogeneous group of companies in terms of business activity, whereas group 2 is not homogeneous.

The financial variables available were: gross revenue, profit, export revenue and research and development expenses (for group 1); and, gross revenue (for group 2). Inlink data to each company website were collected using major commercial search engines at that time: Google, AltaVista, AllTheWeb and MSN Search. Also, the variable *website age* was used based on previous evidence (Vaughan and Thelwall, 2003) showing that inlinks to a website correlated with the age of the site, this is, older sites receives more inlinks. This data were retrieved from the Wayback machine in the Internet Archive (www.archive.org).

Spearman correlations in Table 1 show significant relationships between inlink counts and the three accounting variables. Results suggest that inlinks could be considered as a complementary performance indicator of a company. Vaughan and Wu (2004: 494) suggest that the strong correlation found with research and development expense “could mean that companies that invest more in research and development have a better Web presence and that their sites are more visible and attract more links to them”. No significant correlation was found between inlink count and export revenue. However, export revenue only represents a small fraction (around 8%) of the gross revenue and therefore it seems not to be of relevance as a global performance indicator.

Table 1: Spearman's correlations between business performance measures and inlinks (Vaughan and Wu, 2004)

	Gross revenue	Profit	Research and development expenses
Inlink count	.51	.30	.64
Inlink count / Website age	.50	.30	.63
All correlation coefficients in the table are statistically significant at .01 level.			

In relation to group 2 consisting of heterogeneous companies, there is no significant relationship

between inlink count and gross revenue. As previously mentioned, this group is made up of companies belonging to different industries. Vaughan and Wu (2004: 494) conclude that “link count can be an indicator of business performance only when homogeneous group of companies are being compared”.

In a second paper, Vaughan (2004b) studied the same relationships for the top 100 IT companies in China and the top 51 IT companies in the United States (US). Spearman's test in Table 2 shows significant correlations, lending support to the conclusion raised in the previous study. It is worth noting that the two sets of correlation coefficients are very similar in spite of the remarkable differences between both countries.

Table 2: Spearman's correlations between business performance measures and inlinks (Vaughan, 2004b)

	Inlink count & Gross revenue	Inlink count & Profit	Inlink count / Website age & Gross revenue	Inlink count / Website age & Profit
China	.51	.30	.50	.30
United States	.51	.35	.58	.37
All correlation coefficients in the table are statistically significant at .01 level.				

In a third study (Vaughan, 2004a), all Canadian and United States IT companies were examined. This time, in order to raise more general conclusions, the whole population of the companies in the industry was analysed. Also a new variable, the number of employees, is used to measure the size of the company. This variable is intended to oversee the effect that a larger company will tend to have larger revenue if the rest of the variables remain constant. Apart from confirming previous evidence, the results in Table 3 demonstrate that there is still a significant correlation, even after considering the company size.

Table 3: Correlation between business performance measures and inlinks (Vaughan, 2004a)

	Inlink count & Number of employees	Inlink count & Revenue	Inlink count & Revenue per employee	Inlink count / Website age & Revenue	Inlink count / Website age & Revenue per employee

Canada	.57	.55	.35	.55	.36
United States	.68	.71	.53	.67	.51
All correlation coefficients in the table are statistically significant at .01 level.					

More recently, new research has sought to confirm this evidence by exploring different types of industries.

Romero-Frías and Vaughan (2009a) analysed the top 50 international banks by using two different webometric techniques, inlink analysis and co-link analysis. The banking industry was selected due to its economic significance in the financial crisis context and its high level of internationalization. Two sets of inlink counts (December 2008 and June 2009) and a set of financial variables for the year 2007 (including total assets, total liabilities, total revenue, net income, earnings before tax and return on assets) were taken into account in the study. Spearman's test showed that a majority of the correlation coefficients were significant at the .01 level. Only return on assets was found not significant.

Findings show that inlink counts could be used as a gauge of the banks' financial position and financial performance measures in absolute terms. However, there is no evidence when we refer to relative financial measures, such as return on assets. This could be explained by the absolute nature of the figure “inlink counts”, as it accumulates all links pointing to a webpage since its creation. This is similar to the nature of financial position variables and, somehow, to the financial performance measures as revenue or total income tend to increase over time based on previous performance.

These results are also consistent with the results reported by Romero Frías, Vaughan and Rodríguez Ariza (2009). This study extended previous research by finding evidence of significant correlations between inlink counts and financial variables in several industries in the United States. The study analysed five different industries (commercial banks, construction of buildings, general merchandise store, utilities and mining), as well as the companies in the Dow Jones Industrials. The following economic variables for the years 2005-2007 were collected: number of employees, total assets, net income, total revenue, EBITDA, return on assets (ROA) and return on equity (ROE). Inlink data were retrieved from Yahoo! in January and May 2009. Due to the non normality of the inlink variables and other financial variables, Spearman's test was used. Correlations for the Dow Jones set of companies were not significant, confirming the evidence that only homogeneous companies in terms of activity are comparable (Vaughan and Wu, 2004). Significant correlations were found, to different extents, for all the industries, except for the Construction one. The Store industry had the positive and highest correlation coefficients with all the financial variables, including ROA and ROE. Utilities was the second industry regarding the level of correlation and the

number of variables that were correlated, followed by the Banking and Mining industries.

In comparison to the results of Romero-Frías and Vaughan (2009a), correlation coefficients for the US banking industry are higher than for the global banking industry. This could be explained by the homogeneous competitive conditions that exist in the US market versus the heterogeneous markets where the top international banks operate. For instance, the extent to which the Internet is used for commercial purposes in the different countries could also be an explanatory variable.

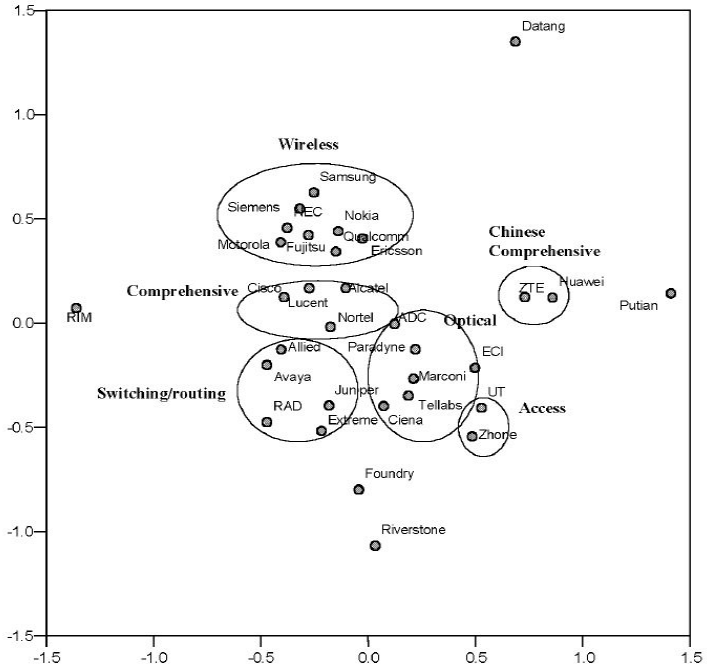
It is worth remarking that correlation does not mean causation. The large number of inlinks that a company's website attracts does not cause the better financial performance. Although it is clear that a positive web image may constitute an intangible asset for a company and therefore can generate future incoming resources, it is more feasible that a bank that is doing well financially is able to maintain a high profile on the Web, maybe through the development of e-commerce practices. As the economy becomes more and more digital and information-based, companies are prompted to monitor if their web presence is in accordance to their financial importance. Web presence measured by inlink count could be an appropriate gauge to evaluate intangible assets related to the Internet.

3.2. Co-link analysis

Web co-link analysis for business information started by focusing on a single industry (Vaughan and You, 2006) or on a specific sector within an industry (Vaughan and You, 2008; Vaughan and You, 2009).

Based on the results of the past studies, Vaughan and You (2006) applied co-link analysis to map business competitive positions of 32 telecommunication companies. The hypothesis under examination is that the number of co-links to the websites of each pair of companies is a measure of the similarity between the two companies. This means that the more co-links the two companies have, the more related they are from the point of view of the sites that link to them. 32 companies were selected according to the following criteria: companies from different parts of the world, from different sectors within the telecommunications industry and top companies in terms of revenue. Co-link data were collected in order to reflect the business relationships in two markets, the global market and the Chinese market. With this purpose, global Yahoo! and Yahoo! China, the top search engine in China at the time of the research, were used respectively to collect the data. The symmetrical co-link matrix obtained was analysed by using multidimensional scaling (MDS), which generated a map showing the relative positions of the companies in the industry. An MDS map depicting the relative positions of companies in the global market is shown as an example in Figure 3.

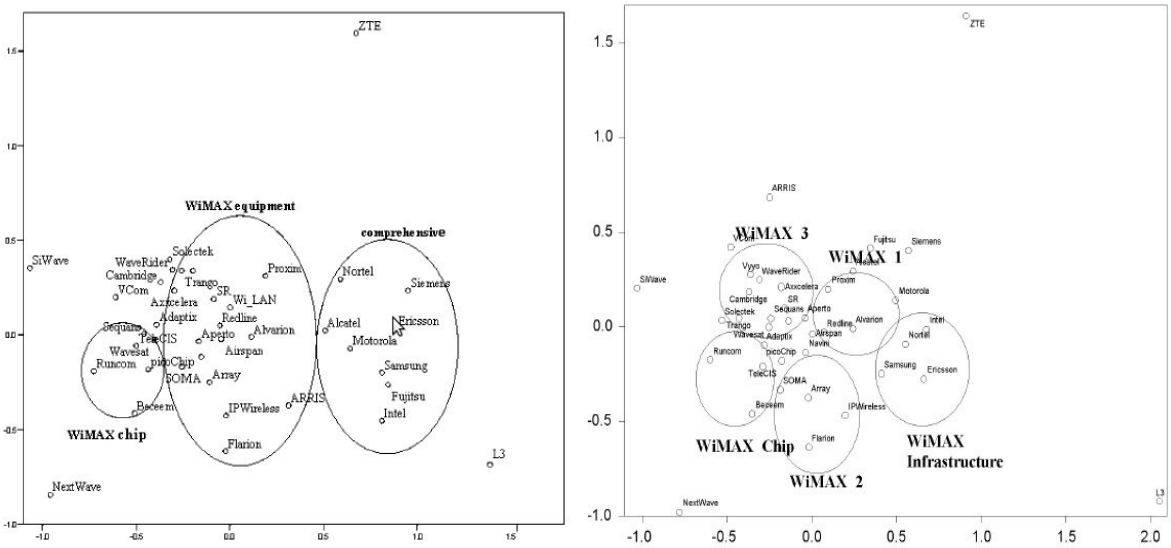
Figure 3: MDS mapping result based on global Yahoo! data (Vaughan and You, 2006: 619)



The companies are clearly clustered into the sectors of the telecommunications industry as labeled in Figure 2. The companies that are not grouped into clusters present specific features which explain their positions. The authors concluded that the maps obtained are consistent with the competition landscape of the industry. Findings suggest that co-link data do contain information about the relationship among companies. Vaughan and You (2006: 618) assert that: “Highly co-linked companies are highly related in their products and the market. Since related companies are competitors (they serve the same market needs), it follows that co-link data can be used to map the competitive position of companies.”

A more recent paper (Vaughan and You, 2008) proposed a method that combines page content mining (keyword) and Web structure mining (co-link data) to achieve a more detailed picture of a particular sector within an industry. The WiMAX sector of the telecommunication industry was chosen specially because this acronym is used only to refer to this particular technology and therefore it avoids the problem of false matches. 39 companies were included in the study and two sets of co-link data were collected (with and without keyword) using the MSN search engine. By adding the keyword WiMAX to the search, only webpages that mention that word are retrieved. This implies removing pages that co-link the two companies for reasons other than the companies' activities in this particular sector. The maps obtained by applying Multidimensional Scaling are shown in Figure 4.

Figure 4: MDS map without the keyword (left) and MDS map with the keyword WiMAX (right) (Vaughan and You, 2008: 438 and 439)



The first map, without the keyword (left), shows clusters of companies in terms of their overall competitive positions in the telecommunication industry. Three main groups are identified: “WiMAX chip”, “WiMAX equipment” and “comprehensive” companies that offer a wide variety of products and services. The second map, with the keyword (right), is able to present companies in five main groups that reflect their competitive positions within the WiMAX sector. This map shows a more detailed analysis on the relationships between telecommunication companies but only from the point of view of the WiMAX services they provide. Authors compared their analysis of the sector with results obtained by an independent market research, finding their conclusions a great match. This study proves that by adding a keyword to the query in the search engine, the information obtained about the relative positions of the companies in the sector is more accurate and meaningful.

The methodology developed in these papers has also been tested and verified in other countries and other industries. Romero-Frías and Vaughan (2009a) analysed the evolution of the top 50 international banks through co-link analysis, finding the existence of 3 main clusters of banks: Asian, European and the so called English-speaking banks that include banks from United States, United Kingdom, Australia and Canada. Vaughan, Tang and Du (2009) studied China’s chemical industry and electronics industry. Finally, Romero-Frías and Vaughan (2009b) extended the use of co-link analysis into the banking industry in the US in order to test the feasibility of combining page content with co-link data to monitor financial crisis.

Parallel to these studies on homogeneous sets of commercial websites, research on heterogenous websites has been carried out to test the triple helix theory on the Web (Stuart and Thelwall, 2006; García-Santiago and de Moya-Anegón, 2009). In this line, Romero-Frías and Vaughan (2010) extended co-link analysis to websites of heterogeneous companies belonging to five stock exchange indexes. The companies selected are the biggest in their respective economies and therefore have a significant web presence. In addition, they are also likely to receive more attention from users, competitors, government and other economic agents. When applying co-link analysis to companies belonging to different industries, the interpretation of co-link as a similarity measure does not stand only for competitive relationships, but for a wide variety of interactions, such as alliances and other linkages between distinct industries. The main conclusion of the study indicates that the degree in which the industrial activity is information centered determines the position of the companies in the MDS maps of the different indexes. The so called information centered industries include mainly IT, media, and financial companies, among others. Co-link analysis could reveal the extent to which certain industries are involved in an ongoing process of business model transformation.

3.3. Content analysis

Qualitative research is a necessary complement to quantitative research because it provides confirming evidence about the relevant nature of the links being analysed. First study on commercial websites was carried out by Vaughan, Gao and Kipp (2006). This study examined three items from a set of webpages: country location of linking webpages, types of sites that created the links and motivations for linking. They analysed 418 links to US companies and 390 links to Canadian companies randomly taken from inlinks to companies studied in Vaughan (2004a). The results conclude that the vast majority of links to business sites are business related, supporting the relevance of link impact analysis for data mining purposes on commercial sites. Regarding motivations for linking, findings show that most types of links came from online directories (22,5%), list of companies (19,6%) and news articles (12,4%). The predominance of inlinks described as directories and list of companies, besides the fact that only 2 out of the 808 links studied pointed to competitors, indicate that co-link could be a fruitful direction to study business competition, as already explained in the previous section.

Content analysis has also been carried out to study the motivations for co-link creation. Vaughan, Kipp and Gao (2007) used content analysis of co-link pages to examine whether co-links had business purposes and why co-links were created. 495 co-link pages to 32 telecommunication companies (the same as used in Vaughan and You, 2006) were classified. 68.5% of co-links were created by commercial sites and 57.6% of co-links pointed to related products or companies. Also, these authors (2007: 447) conclude that “co-links to homepages are more likely to connect highly

related businesses and to show a true business relationship between co-link companies."

4. Concluding ideas and future research

This paper has offered an overview of the webometric methodology to perform quantitative research of the web phenomena. The increasing importance of a hyperlinked society highlights the relevance of this approach to business studies. So far, most of the research on commercial websites has been carried out from an information science perspective, but there is a promising and wide field to explore different types of business issues by using information extracted from the Web. Additional work still needs to be done in analysing different industries and regions and in developing business applications to benefit from the findings up-to-date. It would be useful to monitor the Web periodically to analyse how web variables and economic variables correlate over time. Moreover, explanatory models based on multivariate regression need to be explored.

To conclude, the use of alternative sources to collect web data represent a new horizon for researching in social sciences. In this line, an improvement in the possibilities offered by search engines is to be expected, in particular based on a semantic analysis of internet contents and in the development of more powerful APIs.

5. References

Almind, T. C. and Ingwersen, P. (1997) 'Infometric analyses on the World Wide Web Methodological approaches to 'webometrics'', *Journal of Documentation*, 53(4), pp 404-426.

Bar-Ilan, J. (2004) 'The use of Web search engines in information science research', in Cronin, B. (Ed.), *Annual review of information science and technology*, pp. 231–288, Medford, NJ: Information Today.

Battelle, J. (2005) *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*, London: Portfolio.

Benoît, G. (2002) 'Data mining', in Cronin B. (ed.), *Annual Review of Information Science and Technology*, Vol. 36 (1), pp 265-310, Medford, NJ: Information Today.

Berners-Lee, T. (1999) *Weaving the Web*, San Francisco: Harper.

Björneborn, L. (2004) *Small-world link structures across an academic Web space: A library and information science approach*, Doctoral dissertation, Royal School of Library and Information Science, Copenhagen, Denmark, [online], Available: http://vip.db.dk/lb/phd/phd-thesis.pdf, [consulted 21 Jul 2008].

Björneborn, L., and Ingwersen, P. (2001) 'Perspectives of webometrics', Scientometrics, 50(1), 65–

Björneborn, L., and Ingwersen, P. (2004) 'Toward a basic framework for webometrics', *Journal of the American Society for Information Science and Technology*, 55(14), pp 1216–1227.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual Web search engine', *Computer Networks and ISDN Systems*, 30, pp 1-7.

Evans, M. P. (2007) 'Analysing Google rankings through search engine optimization data', *Internet Research*, 17(1), pp 21-37.

Foot, K. and Schneider, S. (2006) *Web campaigning*, Cambrigde, MA: MIT Press.

García-Santiago, L. and de Moya-Anegón, F. (2009) 'Using co-outlinks to mine heterogeneous networks', *Scientometrics* , 79(3), pp. 681-702.

Garfield, E. (1979) *Citation indexing: Its theory and applications in science, technology and the humanities*, New York: Wiley.

Gomes, B. and Smith, B.T. (2003) Detecting query-specific duplicate documents. U.S. Patent 6,615,209, [online], Available: http://www.patents.com/Detecting-query-specific-duplicate-documents/US6615209/en-US/, [consulted 15 Nov 2009].

Google blog (2008) 'We knew the web was big...', [online], Available: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, [consulted 14 Jan 2009].

Hine, C. (2008) 'Internet Research as an Emergent Practice', in Hesse-Biber, S. N. and Leavy, P. (2008) *Handbook of Emergent Methods*, pp 525-541, New York: The Guilford Press.

Jenkins, H. (2006) *Convergence Culture: Where Old and New Media Collide*, Cambridge, MA: MIT Press.

Kahaner, L. (1996) *Competitive Intelligence – How to Gather, Analyze, and Use Information to Move Your Business to the Top*, 7th ed., New York, NY: Touchstone.

Keen, A. (2007) *The Cult of the Amateur: How the Democratization of the Digital World is Assaulting Our Economy, Our Culture, and Our Values*, New York: Doubleday Currency.

Lewandowski, D., Wahlig, H. and Meyer-Bautor, G. (2006) 'The freshness of web search engine databases', *Journal of Information Science*, 32(2), pp 131–148.

O'Reilly, T. (2005) 'What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software', [online], Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [consulted 17 Jan 2009].

Park, H. W. & Thelwall, M. (2008) 'Web linkage pattern and social structure using politicians' websites in South Korea', *Quality & Quantity*, 42(6), pp 687-697.

Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*, New York, NY: Free Press.

Reid, E. (2003) 'Using Web link analysis to detect and analyze hidden Web communities', in: Vriens, D. (Ed.). *Information and Communications Technology for Competitive Intelligence*, pp 57-84, Hilliard, Ohio: Ideal Group Inc.

Romero Frías, E., Vaughan, L. and Rodríguez Ariza, L. (2009) 'El recuento de enlaces a sitios Web comerciales como indicador de las variables de desempeño y posición financiera de la empresa: estudio empírico de diversos sectores empresariales en Estados Unidos', XV Congreso de la Asociación Española de Contabilidad y Administración de Empresas, Valladolid, Spain, September 23-25, 2009.

Romero-Frías, E. and Vaughan, L. (2009a) 'A Webometric analysis of the global banking industry', 35th EIBA Annual Conference, Valencia, Spain, December 13-15, 2009.

Romero-Frías, E. and Vaughan, L. (2009b) 'Financial Distress of U.S. Banking Industry Viewed through Web Data', 12th International Conference on Scientometrics and Informetrics ISSI 2009, Rio de Janeiro, Brazil, July 14-17, 2009.

Romero-Frías, E. and Vaughan, L. (2010) 'Patterns of Web Linking to Heterogeneous Groups of Companies: The Case of Stock Exchange Indexes', to appear in Aslib Proceedings.

Shaw, D. (2001) 'Playing the links: interactivity and stickiness in .com and 'not.com' websites', *First Monday*, 6(3), [online], Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/837/746 [consulted 10 May 2009].

Sherman, C., and Price, G. (2001) *The invisible Web*, Medford, NJ: Information Today, Inc.

Small, H. (1973) 'Co-citation in the scientific literature: A new measure of the relationship between two documents', *Journal of the American Society for Information Science*, July-August: pp 265–269.

Smith, A. and Thelwall, M. (2002) 'Web Impact Factors for Australasian universities', *Scientometrics*, 54, pp 363-380.

Stuart, D. and Thelwall, M. (2006) 'Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry', *Research Evaluation*, 15(2), pp 97–106.

Tan, B., Foo, S. and Hui, S. C. (2002) 'Web information monitoring for competitive intelligence', *Cybernetics and Systems*, 33(3), pp 225-251.

Technorati (2009) 'State of Blogosphere', [online], Available at: http://technorati.com/blogging/article/state-of-the-blogosphere-introduction/ [consulted 15 Nov

2009].

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

Thelwall, M. (2008a) ‘Extracting accurate and complete results from search engines: Case study Windows Live‘, *Journal of the American Society for Information Science and Technology*, 59(1), pp 38-50.

Thelwall, M. (2008b) ‘How are social network sites embedded in the web? An exploratory link analysis‘, *Cybermetrics*, 12(1), paper 1.

Thelwall, M. (2008c) ‘No place for news in social networking web sites?‘, *Online Information Review*, 32(6), pp 726-744.

Thelwall, M. (2008d) ‘Quantitative comparisons of search engine results‘, *Journal of the American Society for Information Science and Technology*, 59(11), pp 1702-1710.

Thelwall, M. (2008e) ‘Text in social network web sites: A word frequency analysis of Live Spaces‘, *First Monday*, 13(2).

Thelwall, M. (2009) *Introduction to Webometrics. Quantitative Web Research for the Social Sciences*, Morgan & Claypool.

Thelwall, M., Vaughan, L. and Björneborn, L. (2005) ‘Webometrics‘, in Cronin B. (ed.), *Annual review of information science and technology*, 39, pp 81-135, Medford, NJ: Information today.

Thelwall, M. and Wilkinson, D. (2008) ‘A generic lexical URL segmentation framework for counting links, colinks or URLs‘, *Library & Information Science Research*, 30, pp. 515–526.

Turow, J. and Tsui, L. (eds.) (2008) *The Hyperlinked Society: Questioning Connections in the Digital Age*, Ann Arbor: The University of Michigan Press.

Vaughan, L. (2006) ‘Visualizing Linguistic and Cultural Differences Using Web Co-Link Data‘, *Journal of the American Society for Information Science and Technology*, 57(9), pp 1178-1193.

Vaughan, L. (2004a) ‘Exploring website features for business information‘, *Scientometrics*, 61(3), pp 467-477.

Vaughan, L. (2004b) ‘Web hyperlinks reflect business performance—A study of US and Chinese IT companies‘, *Canadian Journal of Information and Library Science*, 28(1), pp 17–31.

Vaughan, L. and Thelwall, M. (2003) ‘Scholarly use of the Web: What are the key inducers of links to journal web sites?‘, *Journal of the American Society for Information Science and Technology*, 54, pp 29-38.

Vaughan, L. and Thelwall, M. (2004) ‘Search engine coverage bias—Evidence and possible causes‘, *Information Processing and Management*, 40(4), pp 693–707.

Vaughan, L. and You, J. (2006) ‘Comparing business competition positions based on Web co-link data—The global market vs. the Chinese market‘, *Scientometrics*, 68(3), pp 611–628.

Vaughan, L. and You, J. (2008) ‘Content assisted web co-link analysis for competitive intelligence‘, *Scientometrics*, 77(3), pp 433-444.

Vaughan, L., and Wu, G. Z. (2004) ‘Links to commercial websites as a source of business information‘, *Scientometrics*, 60(3), pp 487–496.

Vaughan, L., and You, J. (2009), ‘Keyword enhanced Web structure mining for business intelligence‘, *Lecture Notes in Computer Science*, Vol. 4879, pp. 161–168.

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

Vaughan, L., Kipp, M. and Gao, Y. (2007) ‘Are co-linked business web sites really related? A link classification study‘, *Online Information Review*, 31(4), pp 440–450.

Vaughan, L., Tang, J. and Du, J. (2009) ‘Examining the robustness of Web co-link analysis‘, *Online Information Review*, 33(5), pp. 956-972.

Zuccala, A. (2006) ‘Author Cocitation Analysis Is to Intellectual Structure A Web Colink Analysis is to...?‘, *Journal of the American Society for Information Science and Technology*, 57(11), pp 1487-1502.

3 EMPIRICAL RESEARCH

The empirical research is divided in three parts:

1. **Link impact analysis**: exploring the relationships between inlinks received by business Websites and financial variables.
2. **Co-link analysis**: based on the analysis of heterogeneous groups of companies and on the analysis of the financial crisis in the banking industry in U.S.
3. **Combined approach**: including the analysis of the international banking industry based on the two previous approach.

3.1 Link impact analysis

3.1.1 Link impact analysis of different industries in Spain and United Kingdom

The following study has not been published or presented in any conference yet.

Exploring connections between Web inlink data and financial information: A comparison of multiple industries in Spain and the United Kingdom

Abstract

Previous research applied webometric techniques to the study of companies and found that links to commercial Websites contain useful business information. Inlink analysis has shown that hyperlink counts are significantly correlated with financial variables. This research seeks to obtain exploratory evidence about these relations in the European context. Spain and the United Kingdom were selected as they represent two big economies in the European Union and the two languages are among the most commonly used on the Internet. The study included four financial variables, more than what have been explored by previous studies. The study confirmed some findings from the previous studies in the European context. It also found that the Web hyperlink data reflected some unique features of the EU economy. For example, the correlations between the inlink data and the financial data do not change significantly when data from the two countries are merged, reflecting the fact that two countries share a common market. Comparing results from the current study with that from previous studies of other countries such as the U.S., the study also found issues that need further exploration to gain a deeper understanding of the relationship between Web data and financial variables.

1.- Introduction

Two decades after the creation of the World Wide Web, the Internet has become an important space where many people and organizations develop a significant part of their activities all over the world. Castells (2001: 2) refers to the Internet as "a communication medium that allows, for the first time, the communication of many to many, in chosen time, on a global scale". This medium is in a neverending process of transformation due to the ongoing interactions of millions of users in real time. In addition, over the past five years, the Web has undergone relevant changes by the popularisation of the so-called Web 2.0 (O'Reilly, 2005) that has contributed to a dramatic increase in the amount of information available and to a wider variety of users participating in the Internet.

The huge collection of information available makes possible to define the Web as an enormous, unstructured and heterogeneous database that constitutes a unique laboratory to observe social phenomena from different perspectives and by using distinct types of data. Data mining techniques have found a fruitful new space of research by using mainly three types of data: Web content, Web structure and Web usage. In this context, a new discipline, Webometrics (Almind and Ingwersen, 1997; Björneborn and Ingwersen, 2004), has emerged to study quantitatively the Web.

The hyperlink is a key raw material in webometric studies. It can be defined as a reference or navigational element in a document to another section of the same document or to another document. Hyperlinks build the structure of the Web and can be regarded as an endorsement of a target page when created for meaningful purposes. In this paper, we use the concept of *inlink*, which refers to a link coming intoto a Webpage, e.g. page X has an inlink coming from page Z (Björneborn and Ingwersen, 2004). Exploitation of hyperlinks is well illustrated by the functioning of commercial search engines; for instance, Google's search engine dominance derived from the exploitation of *Pagerank* system (Brin and Page, 1998). This system relies on the idea that a link pointing to a Webpage means a vote to that Webpage or document. Nevertheless, this reasoning is far from being new. It resembles clearly the Garfield's proposal (1979) to use bibliographic citations as the basis to rank scientific production, a technique that is still used nowadays to evaluate the quality of academic research. In general, we could say that Webometrics adapts Bibliometrics analysis to the study of the Web (Thelwall, Vaughan and Björneborn, 2005), although it is progressively developing genuine techniques.

The application of webometric techniques to commercial Websites is in a developing phase. Competitive Intelligence (CI) is one of the areas where this research has been more successful and promising (Tan *et al*., 2002; Reid, 2003). Kahaner (1996) considers that CI consists of a systematic plan to obtain and analyse information about competitors and general trends in the industry. The abundance of digital information available in the Internet opens new possibilities for companies and other economic agents to analyse data publicly available and generate valuable knowledge. For instance, a recent study (Choi and Varian, 2009), carried out by Google

researchers, used data from the Google Trends service to anticipate consumer behaviour in several industries.

An especially productive area of research has been based on *link impact analysis* (Thelwall, 2004), which is based on the number of web pages that link to a collection of pages or sites in order to assess their impact. Vaughan and colleagues used this technique to explore quantitative relationships between inlink counts to commercial Websites and business performance variables. Vaughan and Wu (2004) tested the hypothesis in two groups of Chinese companies: one homogeneous group (in terms of industry) made of China's top 100 Information Technology (IT), and one heterogeneous group made of top 100 privately owned companies. Spearman correlation tests showed significant relationships between inlink counts and a set of financial variables (i.e. gross revenue, profit, and research and development expenses). However, these relations were not significant when companies from different industries were analyzed together. Vaughan (2004a; 2004b) reinforced this evidence by studying IT industry in China, USA and Canada. It is worth noting that the two sets of correlation coefficients for China and USA were very similar in spite of the remarkable differences between the two countries.

Romero-Frías and Vaughan (2009a) analysed the top 50 international banks. Spearman's test showed that a majority of the correlation coefficients between inlinks and a set of financial variables (i.e. total assets total revenue, net income and earnings before tax) were significant at the .01 level. Only return on assets was found not to be statistically significant. These results are also consistent with Romero Frías, Vaughan and Rodríguez Ariza (2009), which extended previous research by finding evidence of significant correlations between inlink counts and financial variables in several industries in the United States. The study analysed five different industries (commercial banks, construction of buildings, general merchandise store, utilities and mining), as well as the companies in the Dow Jones Industrials.

Findings showed that inlink counts could be used as an indicator of the business financial position and financial performance measures in absolute terms. However, there is no evidence of this when we refer to relative financial measures, such as return on assets. This could be explain by the nature of the variable “inlink” that represents the number of links pointing to a particular Webpage or Website since its creation. The functions offered by search engines, at this moment, do not allow us to retrieve inlinks created during a specific period of time. Somehow, this is similar to the nature of financial position variables that accumulates over time. Financial performance measures as revenue or total income tend also to increase over time based on previous performance.

The purpose of this research is to obtain exploratory evidence about the relations between the

number of links pointing to a company website and the financial variables of that company in the European context. European companies constitute a challenge because different economies, languages and cultural backgrounds are in play within a single market. Also this research is of especial interest because unlike the US, financial data for private companies are available for European countries.

2.- Methodology

2.1.- Selection of companies to study

Five industries with very different activities and Web presence were selected. Spain and United Kingdom are included in the study because they represent two big economies in the European Union and two of the most influential languages in the Internet. Information about the companies in the study was gathered in October 2009 from Amadeus (http://www.bvdep.com/en/amadeus.html), a reliable database containing financial information on European public and private companies. Cases retrieved from the database were filtered out in order to match the following criteria:

- The last year of information available is 2007.
- All the companies have, at least, information about total assets and Profit (loss) after tax.
- Companies have a website reported in the database.

Table 1 shows the industries in the study, their NAICS code for industrial classification and the number of companies for each country.

Table 1. Country and Industry Distribution of the Companies

Industries (NAICS 2007 code)	Spain	United Kingdom	Total
Construction (23)	93	103	196
Accommodation and Food Services (72)	70	180	250
Utilities (221)	56	60	116
Publishing Industries (except Internet) (511)	72	201	273
Telecommunications (517)	26	71	97
Total	317	615	932

The Website address of each company was collected from the Amadeus database and then

manually checked to ensure its correctness. Companies that shared a website with another company were deleted unless one of those companies could be clearly identified as the company delivering the services or the company that was significantly bigger in terms of total assets. For the companies that have alternative URLs in the form of alias or redirect, we checked each URL and select the one that has more inlinks.

2.2. Collecting financial data

As previously mentioned, financial data were collected from the Amadeus database in October 2009. A set of relevant financial variables were retrieved: total assets, profit (loss) after tax, operating revenue / turnover, and intangible fixed assets. The number of cases available for each variable is shown in tables 2, 4 and 5.

2.3. Collecting Web link data

Of the three major search engines, Google, Yahoo! and MSN Bing, only Yahoo! could be used for data collection for the study. Google's inlink search only returns a sample of all inlinks that the Google database records (Google, 2009). Moreover, Google operators are more limited for this purpose and cannot filter out internal inlinks (for example links created to the "homepage" for navigational purposes) as the query term 'link" cannot be combined with any other query terms (Google, 2006). In other words, it cannot report the external inlink counts that the study needed. The MSN search engine used to have inlink search functions but the service was turned off around March 2007 (Live Search, 2007). At the time of data collection, November 2009, Yahoo! is the only option available for collection of the data required for the study.

Because search engines of different countries may have databases that favor Websites of the host countries (Vaughan and Thelwall, 2004), we considered using Yahoo! Spain and Yahoo! UK respectively for companies of each country. However, tests of these two versions of Yahoo! showed that they returned the same inlink search results and therefore Yahoo! UK interface was used to collect the data.

Yahoo! has two inlink search query terms, link and linkdomain. The "link" query term finds links to a particular page (e.g. link:http://www.abc.com finds links to the homepage of www.abc.com) while the linkdomain query term retrieves all links that point to all pages of a particular Website or domain including the homepage. We used the linkdomain query term for data collection because all link are of relevance to the study. The query syntax for the data collection, using the hypothetical URLs of www.abc.com is the following: linkdomain:abc.com –site:abc.com. We truncated the www portion of the URLs in the queries to capture all links to all subdomains such as

mail.abc.com. The “-site:abc.com” part of the query is to filter out internal links coming from within the domain of abc.com itself.

3.- Results

3.1.- Correlations between Web inlink and financial data by industry and country

The frequency distributions of the inlink variable and the financial variables are very skewed, so the Spearman correlation test rather than the Pearson correlation test was used to explore the relationships between variables.

The correlation coefficients between the inlink counts and the financial data for companies by industry and by country are shown in Table 2. The columns below the label “industries” show correlations for companies in the same industry without distinguishing the country, whereas the columns below the label “country” show correlations for companies in the same country irrespective of their industry. As shown by previous research (Vaughan and Wu, 2004; Romero Frías, Vaughan and Rodríguez Ariza, 2009), correlations are likely to be significant when a single industry is considered. This is explained by the different business models and Web presence patterns which are dominant in each industry. Therefore, it is expected that correlations are more significant when we talk in term of homogeneous companies. Table 2 shows that although correlations are significant in both set of columns, the coefficients are higher, for a majority of variables, when considered at industry level.

The correlation for the variable “intangible fixed assets” is either not significant (construction and utilities) or low. Nevertheless, the publishing and telecommunications industries, which are very focused on information and technology, show the highest significant correlation coefficients, .178 ($p<0.01$) and .250 ($p<0.05$) respectively. This could suggest the weaknesses in the treatment for intangible assets recognition and measurement, one of the most challenging issues in the Accounting field. However, although Web presence is an important intangible asset for a company and could also be reflecting other intangible assets, not all of them need to be related in some ways to the Web.

The highly significant correlation coefficients by industries indicates that despite the difference between countries there is a certain level of homogeneity within the same industry irrespective of the place were the company is established. This homogeneity could be a result of the process of economic globalization, that is especially acute in large and very large companies such as the ones in the study (although the majority of them are not public companies). Membership to the European Union is also a relevant feature because companies compete in a single European market. However, it is worth noting that Spain belongs to the Eurozone, but not United Kingdom.

The variable “total assets” seems to be the best gauge for inlink counts in the five industries, with the lowest correlation coefficient for the accommodation industry (.369***). It is followed by the variable “operating revenue”. “Profit (loss) after tax” is clearly significant at p=.001 for the construction and publishing industries, at p=.05 for accommodation and utilities industries, and not significant for telecommunications. The lower correlation for the “profit (loss) after tax” variable in comparison with other variables could be explained by the nature of this variable that can be either positive or negative, whereas the others can only have a minimum of zero. The profit is one of the best indicators of business performance and therefore is more contingent upon the financial evolution of the year. Although this is also the case of the operating revenue, this magnitude remains more stable, or could even increase, period after period, even when the profits of the company decrease or are negative due to a high level of expenses.

Table 2. Correlation between inlink data and financial data in terms of industry and country

Spearman's rho correlation with number of inlinks	Industries					Country	
	Construction	Accomm. Food	Utilities	Publishing	Telecommu-nications	Spain	United Kingdom
Operating Revenue / Turnover	,513***	,260***	,362***	,414***	,367***	,338***	,262***
N	185	235	112	265	95	314	578
Profit (loss) after tax	,305***	,143*	,219*	,267***	,090	,129*	,198***
N	196	250	116	273	97	317	615
Total assets	,470***	,399***	,488***	,369***	,407***	,224***	,351***
N	196	250	116	273	97	317	615
Intangible fixed assets	,140	,153*	,122	,178**	,250*	,317***	,221***
N	181	245	114	265	91	313	583

In order to ascertain if there are significant differences between the variables in each country, Mann-Whitney tests were carried out for companies in each industry. Table 3 reports asymptotic significance (2-tailed) of the statistics.

Table 3. Significance of the Mann-Whitney Tests

	Inlinks	Operating Revenue / Turnover	Profit (loss) after tax	Total assets	Intangible fixed assets
Construction	,001	,001	,818	,100	,000
Accommodation and Food Services	,727	,015	,046	,453	,000
Utilities	,012	,000	,008	,001	,005
Publishing	,004	,825	,986	,197	,070
Telecommunications	,208	,081	,195	,005	,081

These statistics show that there are significant differences between the variables depending on the country. For instance, the inlink count in Spain and the United Kingdom are significantly different for the companies in the Construction, Utilities and Publishing industries, but not in the Accommodation and Food Services and Telecommunications ones. In the next section we examine the correlations in the different industries in Spain and in the United Kingdom separately.

3.2.- Correlation between Web inlink and financial data by industries in Spain and United Kingdom

As already mentioned, we expected that the level of correlation is higher in each specific industry within the country than when all the industries in the country are considered simultaneously (Table 2). The results in table 4 show that this is true only in the case of the construction industry and in the other industries only for certain variables. Total assets is the variable that correlated more significantly with inlinks improving in all the cases the correlation coefficients obtained in table 2 for Spain.

When we compared results in Table 4 with the correlations by industries in Table 2, we find that, in general, only the construction and the telecommunications industries show higher correlations when analyzed at the Spanish level.

Table 4. Correlations between inlink data and financial data by industry in Spain.

Spearman's rho correlation with number of inlinks	Spain				
	Construction	Accomm. Food	Utilities	Publishing	Telecommu-nications
Operating Revenue / Turnover	,554***	,255*	,207	,374**	,481*
N	91	69	56	72	26
Profit (loss) after tax	,343**	,028	,130	,341**	-,022
N	93	70	56	72	26
Total assets	,518***	,317**	,439**	,275*	,612**
N	93	70	56	72	26
Intangible fixed assets	,345**	,324**	,311*	,052	,486*
N	90	69	56	72	26

***. Correlation is significant at the 0.001 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

In Table 5 we have the results for the different industries in the United Kingdom. Results at country level improve for all the variables, except intangible assets, when we compared to the results of United Kingdom in Table 2. This indicates that homogeneous companies are better candidates to find highly significant correlation coefficients.

When compared to industries in Table 2, the analysis at UK level is more significant, in general terms, for the construction and accommodation industries, very similar for publishing and slightly worse for utilities and communications. However, this analysis depends on the variable we consider as there is no clear pattern.

As in the Spanish case, total assets is the variable that correlates more significantly with inlinks.

Table 5. Correlations between inlink data and financial data by industry in the United Kingdom.

Spearman's rho correlation with number of inlinks Sig. (2-tailed)	United Kingdom				
	Construction	Accomm. Food	Utilities	Publishing	Telecommu-nications
Operating Revenue / Turnover	,396***	,277***	,318*	,413***	,341**
N	94	166	56	193	69
Profit (loss) after tax	,301**	,180*	,158	,240**	,190
N	103	180	60	201	71
Total assets	,524***	,417***	,467***	,418***	,398**
N	103	180	60	201	71
Intangible fixed assets	,280**	,113	,121	,184*	,179
N	91	176	58	193	65
***. Correlation is significant at the 0.001 level (2-tailed).					
**. Correlation is significant at the 0.01 level (2-tailed).					
*. Correlation is significant at the 0.05 level (2-tailed).					

Table 6 indicates which country presents a higher significant correlation for each industry and variable. It is worth noting that in general the accommodation, utilities and publishing industries correlate better for companies in the UK, whereas construction and telecommunications for Spanish companies. Also, the variable “total assets” correlates more significantly with inlink counts for UK companies.

Table 6.- Which country has the higher correlation coefficient

Comparison between Spain (ES) and United Kingdom (UK)	Construction	Accomm. Food	Utilities	Publishing	Telecommu-nications
Operating Revenue / Turnover	ES	UK	UK	UK	ES
Profit (loss) after tax	ES	UK	-	ES	-
Total assets	UK	UK	UK	UK	ES
Intangible fixed assets	ES	ES	ES	UK	ES

4.- Discussion

This is the first study of this type for multiple industries, in two different countries and including private companies. These features seem to complicate very much the analysis of the data as no clear overall patterns are found in the results.

First it is important to underline that when the different industries in the same country are aggregated correlations are still significant, although they are lower than when the industries are considered separately. This suggests that larger studies need to be done in particular economies to explore the extent of significant correlations when heterogeneous companies are considered.

Second it is relevant to note that, when the analysis by industry in Table 2 is split into countries, the results do not improve clearly. They are even less significance in some industries and variables. There is evidence that correlation levels are similar in different countries, e.g. in the case of US and China (Vaughan, 2004b) and US and Canada (2004a). This is even more justified in this case as both countries share a single market as members of the European Union. Therefore, this suggests that in some cases economic regions could be more useful in the analysis than countries.

If we compared the results of this study with the results of Romero Frías, Vaughan and Rodríguez Ariza (2009) regarding multiple US industries we note that the behaviour of the two industries included in both works is very dissimilar. The construction industry in the US economy does not show significant correlations with any of the variables, whereas in this study it is one of the industries with the highest correlation coefficients, especially in the Spanish case. In fact, this is a relevant reflection of the importance of the construction industry in Spain , being one of the economic engines of the economy for years. Also, the utilities industry has a much lower level of correlation in the Spain and UK than in the US. For instance, the correlation coefficients between inlinks and total assets are .439 in Spain, .467 in UK and .794 in the U.S.

The study faces some limitations derived from the nature of the web information available. It is not possible to collect inlink data from search engines for previous periods of time. These means that this paper has to use inlink data collected in 2009 whereas the more updated financial data available are of the year 2007. Longitudinal studies can solve this problem in the next years. Also, it is important to note that the different groups of industries selected could be segmented into more specific and homogeneous groups. For example, NAICS code 511, Publishing Industries, includes book publishers, press and software publishers. However this would reduce very much the size of the sample in the study.

Finally, we would like to emphasize that correlation does not mean causation. The number of inlinks that a company's Website attracted does not cause or explain the financial performance, although it represents an intangible asset (i.e. related to the business reputation) that will generate

future resources for the company. Although we cannot establish a causal relationship, the correlation found is still very useful information to elaborate prediction models of one variable based on the other.

4.- Conclusions and future research

The original purpose of this study was to extend evidence on the relationships between web structure variables, inlinks, and financial variables into the European context. Evidence was found that support previous research but also that contradicts previous findings and open new opportunities of research.

The significant correlations found between inlink counts and financial variables show that Web structure data can provide useful business information. This information can be used to provide new insight about valuation and measurement of intangible assets related to the Web presence of a company. The correlation could be used to develop a regression model that predicts what a company's Web position should be based on their financial data. Comparing the predicted Web position with the real one would reveal those that underperform on the Web. Further analysis could then be done to find out why and how to improve.

References

Almind, T. C. and Ingwersen, P. (1997). Informetric analyses on the World Wide Web Methodological approaches to 'webometrics', *Journal of Documentation*, 53(4), 404-426.

Björneborn, L. and Ingwersen, P. (2004). Toward a basic framework for webometrics, *Journal of the American Society for Information Science and Technology*, 55(14), 1216–1227.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine, *Computer Networks and ISDN Systems*, 30, 1-7.

Castells, M. (2001) *The Internet Galaxy. Reflections on the Internet, Business and Society*. Oxford: Oxford University Press.

Choi, H. y Varian, H., (2009). *Predicting the Present with Google Trends*. Available: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf [consulted 10 May 2009].

Garfield, E. (1979). *Citation indexing: Its theory and applications in science, technology and the humanities.* New York: Wiley.

Google (2006). *Google SOAP Search API Reference*. Retrieved June 3, 2009, from http://www.google.com/apis/reference.html#2_2

Google (2009). *Links to your site*. Retrieved June 3, 2009, from http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=55281

Kahaner, L. (1996). *Competitive Intelligence – How to Gather, Analyze, and Use Information to Move Your Business to the Top*, 7th ed., New York: Touchstone.

Live Search (2007). *We are flattered, but...* Retrieved June 3, 2009, from http://www.bing.com/community/blogs/search/archive/2007/03/28/we-are-flattered-but.aspx

O'Reilly, T. (2005) 'What is Web 2.0. Design Patterns and Business Models for the Next Generation of Software', [online], Available: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html [consulted 17 Jan 2009].

Reid, E. (2003). Using Web link analysis to detect and analyze hidden Web communities. In D. Vriens (ed.), *Information and Communications Technology for Competitive Intelligence*, Ohio: Ideal Group Inc, 57-84.

Romero-Frías, E. and Vaughan, L. (2009a) 'A Webometric analysis of the global banking industry', 35th EIBA Annual Conference, Valencia, Spain, December 13-15, 2009.

Romero Frías, E., Vaughan, L. and Rodríguez Ariza, L. (2009). El recuento de enlaces a sitios Web comerciales como indicador de las variables de desempeño y posición financiera de la empresa: estudio empírico de diversos sectores empresariales en Estados Unidos, *XV Congreso de la Asociación Española de Contabilidad y Administración de Empresas*, Valladolid, September 23-25, 2009.

Tan, B., Foo, S. and Hui, S. C. (2002). Web information monitoring for competitive intelligence. *Cybernetics and Systems*, 33(3), 225-251.

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

Thelwall, M., Vaughan, L. and Björneborn, L. (2005). Webometrics. In B. Cronin (ed.), *Annual review of information science and technology*, Medford, NJ: Information today, 39, 81-135.

Vaughan, L. (2004a). Exploring website features for business information. *Scientometrics*, 61(3), 467-477.

Vaughan, L. (2004b). Web hyperlinks reflect business performance—A study of US and Chinese IT companies. *Canadian Journal of Information and Library Science*, 28(1), 17–31.

Vaughan, L. and Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*, 40(4), 693-707.

Vaughan, L., and Wu, G. Z. (2004). Links to commercial websites as a source of business information. *Scientometrics*, 60(3), 487–496.

3.1.2 Link impact analysis of different industries in the United States.

The following study has been presented as a conference paper. This is the reference:

> ROMERO FRÍAS, E.; VAUGHAN, L.; RODRÍGUEZ ARIZA, L. (2009) "Inlink counts to commercial Websites as an indicator of financial performance and financial position variables of companies: empiric evidence in different US industries" (in Spanish). In *Proceedings of the Conference of the Spanish Association of Accounting and Business Management*, Valladolid (Spain), September 23-25, 2009. ISBN: 978-84-96648-31-9.
>
> Available at http://www.aeca.es/pub/on_line/comunicaciones_xvcongresoaeca/cd/132g.pdf

This study was originally written in Spanish; here we include a brief summary with the most important results and conclusions.

1.- Introduction

The introduction to this study is analogous to the introduction in the previous paper (section 3.1.1).

2.- Method

Here we analyzed 5 different industries in the United States, plus the companies in the Dow Jones Industrials. United States was the country selected due to the following reasons:

- The level of Web usage by companies and the population in general is very high.
- It is the most important economy in the world and their companies are likely to be large enough to have a significant presence in the Web.
- Companies belonging to the IT industry have been studied in previous studies (Vaughan, 2004a y 2004b) and therefore the results can be compared.

All the companies are listed. The industries included are: Commercial banks, Construction of Buildings, General Merchandise Stores, Mining, and Utilities. Table 1 describes the sample of companies in the study. Business information was retrieved from the Mergent database (www.mergent.com). The sample is made of 100 companies in each industry. When the number of companies in the industry is lower, we take the entire population. Some companies are excluded due to the following reasons:

- Companies do not have a Web site.
- Financial data for the year 2007 are not available.
- Different companies belonging to the same group share the same Web site.

The column with the label “valid companies” indicates the number of companies included in the statistical tests.

Table 1. Sample of companies

Industries (NAICS code)	Total n. of companies	Active companies	Sample size	Excluded companies (no Website)	Excluded companies (no financial information)	Valid companies	% of active companies
Commercial banks (52211)	1.099	581	100	6	-	94	16 %
Construction of Buildings (236)	88	57	57	16	17	24	42 %
General Merchandise Stores (452)	71	35	35	6	5	24	69 %
Utilities (221)	426	275	100	15 + 3 (sharing URL)	7	75	27 %
Mining (21)	897	671	100	21	26	53	8 %
Dow Jones Industrials	-	30	30	-	-	30	100 %

The composition of Dow Jones Industrials is at December 4 2008 (http://www.djindexes.com/mdsidx/?event=showAverages). The financial variables taken were: net income, total revenue, total assets, total liabilities, number of employees, ROA (Return on assets) and ROE (Return on equity) for the years 2005, 2006 and 2007. Table 2 shows the information regarding the collection of inlink data.

Table2. Collection of inlink data

Queries (Yahoo!)	linkdomain:abc.com -site:abc.com	linkdomain:abc.com -site:abc.com	link:http://www.abc.com –site:abc.com
Method	Web service	Yahoo! API	Yahoo! API
Industries (NAICS code)	Dates (collected from Canada)		
Commercial banks (52211)	30-01-2009	06-05-2009	06-05-2009
Construction of Buildings (236)	02-02-2009	06-05-2009	06-05-2009
General Merchandise Stores (452)	02-02-2009	06-05-2009	06-05-2009
Utilities (221)	01-02-2009	06-05-2009	06-05-2009
Mining (21)	05-02-2009	06-05-2009	06-05-2009
Dow Jones Industrials	08-02-2009	Not available	Not available

3.- Results

The frequency distributions of the inlink variable and the financial variables are very skewed, so the Spearman correlation test rather than the Pearson correlation test was used to explore the relationships between variables.

The results obtained with the inlink count collected at the beginning of the year and using the operator “linkdomain” are the highest. Therefore, we use this inlink count as a reference. Table 3 shows how the correlation between the different inlink counts collected are very high.

Table 3. Correlations between inlink counts at the beginning of 2009 (linkdomain) and in May 2009 (linkdomain and link).

	May 2009 "linkdomain"	May 2009 "link"
Commercial banks (52211)	,964**	,903**
Construction of Buildings (236)	,923**	,890**
Mining (452)	,940**	,949**
General Merchandise Stores (221)	,988**	,977**
Utilities (21)	,982**	,973**

The reference for the correlations is the inlink count taken at the beginning of 2009 using the operator "linkdomain".

** Correlation is significant at the 0.01 level (2-tailed).

Table 4 show the correlation coefficients between inlink counts and financial variables. For each industry the number 1st indicates the inlink count at the beginning of 2009 (using "linkdomain") and the 2nd, the inlink count in May 2009. We take the 1st inlink count and the financial information for the year 2007 as the reference for the analysis.

Tabla 4.4. Spearman's correlation coefficients for the different industries

		Commercial banks		Construction of Buildings		Mining		General Merchandise Stores		Utilities		Dow Jones
	year	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st
Employees	2007	,741**	,699**	,292	,335	,522**	,559**	,845**	,868**	,855**	,847**	,239
Total assets	2007	.736**	.696**	.129	.136	.503**	.534**	.868**	.884**	.837**	.849**	.206
	2006	.719**	.678**	.165	.114	.495**	.498**	.861**	.878**	.794**	.807**	.201
	2005	,744**	,708**	,255	,189	,457**	,532**	,857**	,873**	,759**	,763**	,191
Net income	2007	.627**	.623**	.099	.134	.106	.175	.941**	.929**	.794**	.804**	.319
	2006	.682**	.663**	.218	.140	.116	.196	.897**	.889**	.711**	.709**	.133
	2005	,719**	,684**	,102	,069	,114	,188	,921**	,918**	,576**	,553**	,102
Total revenue	2007	.745**	.707**	.240	.242	.525**	.530**	.886**	.908**	.801**	.809**	.114
	2006	.730**	.684**	.181	.161	.466**	.489**	.863**	.885**	.786**	.790**	.093
	2005	,750**	,711**	,224	,141	,471**	,514**	,839**	,845**	,767**	,771**	,100
EBITDA	2007	-	-	.075	.075	-	-	.944**	.935**	.820**	.826**	.385*
	2006	-	-	.246	.207	-	-	.943**	.939**	.737**	.731**	.379
	2005	-	-	,271	,194	-	-	,901**	,894**	,679**	,662**	,255
ROA	2007	.134	.183	.133	.181	.237	.274	.560**	.476*	.389**	.404**	.123
	2006	.278*	.304**	.074	.005	.101	.197	.570**	.524*	.344**	.316**	-.015
	2005	,242*	,246*	,197	,165	-,114	-,027	,616**	,590**	,194	,166	,031
ROE	2007	.154	.174	.344	.375	.227	.163	.649**	.573**	.050	.032	.023
	2006	.298**	.290**	.174	.081	.236	.206	.671**	.622**	.179	.137	-.119
	2005	,221*	,207	,197	,132	-,002	,015	,626**	,567**	,049	,011	-,048

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Table 5 shows the ranking of each industry in terms of the correlation coefficient (being 1 the industry with the highest coefficient). A hyphen appears when the correlation is not significant. N/A means that the data were not available.

Table 5. Ranking of each industry in terms of the correlation coefficient

	Employees	Assets	Net Income	Total revenue	EBITDA	ROA	ROE
Commercial banks	3	3	3	3	N/A	-	-
Construction of Buildings	-	-	-	-	-	-	-
Mining	4	4	-	4	N/A	-	-
General Merchandise Stores	2	1	1	1	1	1	1
Utilities	1	2	2	2	2	2	-

The results for each industry show the following:

- **Commercial banks**: significant correlations with the main financial variables (in absolute terms) for the year 2007: total assets, 0,736; net income, 0,627; total revenue, 0,745. No significant correlations with ROA and ROE. Inlink counts seem to be a good gauge for business dimension, but not for financial performance.
- **Construction of Buildings**: no significant correlations.
- **Mining**: significant correlations with some financial variables (in absolute terms): total assets, 0,503; total revenue, 0,525. No significant correlations with ROA and ROE.
- **General Merchandise Stores**: significant correlations with all the variables, also with ROA 0,56 and ROE 0,649. The companies in this industry receive more inlinks than companies in any other industry. Its activity is very user-oriented and this could explain the strong presence in the Web.
- **Utilities**: significant correlations with some financial variables: total assets, net income, total revenue, EBITDA and ROA (0,389).
- **Dow Jones Industrials**: In line with Vaughan and Wu (2004), there is not significant correlations in heterogeneous groups of companies, in terms of industry.

4.- Conclusions

The main conclusions are the following:

- In most of the industries, there is a linkage between inlink counts and the financial performance and financial position variables (in absolute terms). Correlation coefficients are very high in the “General Merchandise Stores” industry, which is very user-oriented.
- Significant correlation coefficients are observed in homogeneous groups of companies, not in the Dow Jones Industrials.
- Table 5 shows that the higher the correlation coefficients, the higher the number of financial variables that correlate significantly.
- The number of employees seems to be a good indicator of business size in most of the industries.
- The coefficients are higher than in previous studies (Vaughan, 2004a y 2004b; Vaughan y Wu, 2004). This could be explained by the strong development of the Web in the last 5 years, especially due to the popularization of Web 2.0 tools.

3.1.3 Multivariate regression analysis of the number of inlinks received by business Websites in Spain and United Kingdom

The following study has not been published or presented in any conference yet.

This study was originally written in Spanish; here we include a brief summary with the most important results and conclusions.

1.- Introduction

This study goes beyond simple correlation analysis and uses a mathematical modeling approach that allows the analysis of multiple explanatory variables simultaneously. The modeling approach also allow us to examine another important variable, the country (Spain vs United Kingdom). A previous paper (Vaughan y Thelwall, 2005) used a multiple regression model to explain hyperlink patterns in Canadian university Websites.

2.- Method

This study uses the same data as the study in section 3.1.1.

The dependent variable of the model is the number of inlinks received by a business Web site. The independent variables are total assets, intangible fixed assets, turnover, profit (loss) after tax and a dummy variable for the country of the company (United Kingdom, 0 and Spain, 1). The financial variables selected represent some of the most relevant indicators of the financial position and performance of the company.

Table 1 shows the assumptions for the regression model. We include diagrams in the Appendix to test the assumptions.

Table 1. Assumptions for the regression analysis

Assumptions	Meaning	Tests / Diagrams	Features
Linearity	Linear relationship of the residuals of the dependent variable and each of the independent variables.	Partial plots.	The points are distributed in a linear pattern.
Homoscedasticity	Constant variance of the residuals.	Plot of standardized residuals against standardized predicted values.	The points are randomly and evenly dispersed throughout the plot.
Normality	The residuals in the model are normally distributed.	Histogram of the standardized residuals. P-P plots of normally distributed residuals.	The distribution of the residuals of the model should approximate a normal distribution. In the P-P plot the points should be over the straight line.

Multicollinearity	A situation in which two or more variables are very closely linearly related.	Tolerance	Close to 1. Threshold 0,1.
		VIF	Close to 1. Threshold 10.

SPSS v17 was used for the statistical analysis. The method used for the multiple regression analysis was *Stepwise*, although the *Enter* method was also calculated to confirm the results obtained. The theoretical model we want to test is in table 2.

Table 2. Theoretical model

Equation	$Y_1 = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5$
Dependent variable	Y_1: Inlinks
Independent variables	X_1: Total assets X_2: Intangible fixed assets X_3: Turnover X_4: Profit (loss) after tax X_5: Country: United Kingdom (0), Spain (1)

The hypothesis underlying the model is that the number of inlinks a business Web site receives can be explained by the financial position of the company (as measured by total assets and intangible fixed assets), the financial performance of the company (as measured by turnover and profit after tax) and the country origin of the company (dummy variable). The hypothesis are, in detail:

- Hypothesis 1: There is a positive and significant relationship between the “Total assets” variable and the number of inlinks that a business Website receives.
- Hypothesis 2: There is a positive and significant relationship between the “Intangible fixed assets” variable and the number of inlinks that a business Website receives.
- Hypothesis 3: There is a positive and significant relationship between the “Turnover” variable and the number of inlinks that a business Website receives.
- Hypothesis 4: There is a positive and significant relationship between the “Profit after tax” variable and the number of inlinks that a business Website receives.
- Hypothesis 5: There is a significant relationship between the “Country” variable and the number of inlinks that a business Website receives.

The aforementioned model is examined in each of the 5 industries.

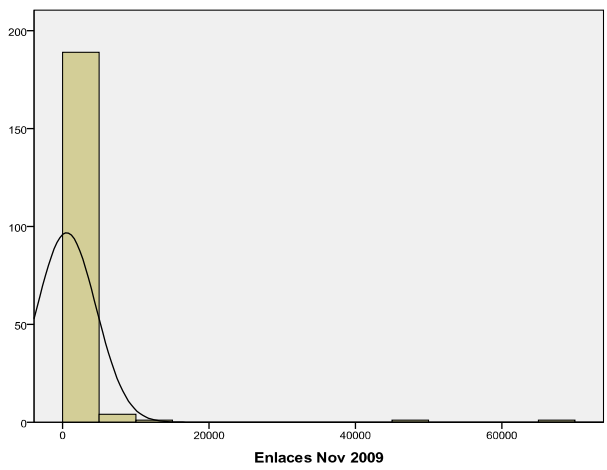
When we calculate the models for the first time, we observed that the residuals are non-normally distributed. Also the distribution of inlink counts for each industry (dependent variable) is very skewed (see table 3 and figure 1). Therefore we decided to apply a log transformation to the data.

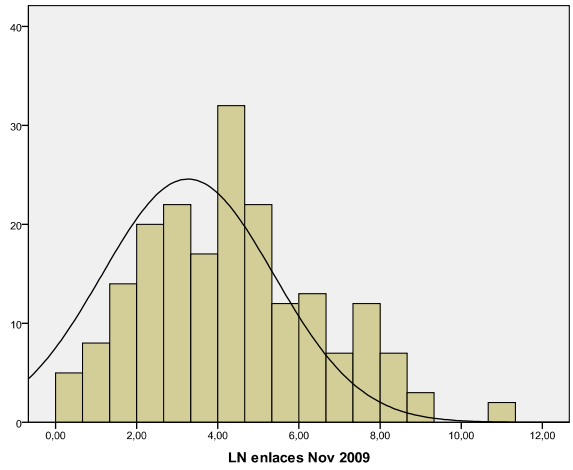
Table 3. Descriptive statistics of the variable “Inlinks” (before and after the transformation)

	Construction		Accommodation and Food		Utilities		Publishing		Telecommunications	
	No trans.	Trans.	No trans.	Trans.	No trans.	Trans.	No trans.	Trans.	No trans.	Trans.
N: Valids (Missing)	196 [ES =93; UK =103]		249 [ES =70; UK =179]		113 [ES =56; UK =57]		273 [ES =72; UK =201]		97 [ES =26; UK =71]	
Mean	1104,75	4,33	7074,55	6,04	7141,65	5,93	275256,92	8,54	55487,46	6,25
Median	62,00	4,14	414,00	6,03	336,00	5,82	3990,00	8,29	235,00	5,46
Standard deviation	5779,34	2,21	58125,68	2,19	29543,36	2,32	918548,21	3,29	338445,60	2,81
Skewness	9,52	,38	14,44	,053	6,56	,55	5,49	,14	8,74	,86
Standard error of skewness	,174	,174	,154	,154	,23	,23	,147	,147	,245	,245
Kurtosis	96,79	-,075	218,35	,36	48,23	-,065	40,03	-,54	80,88	,50
Standard error of kurtosis	,346	,346	,307	,307	,451	,451	,294	,294	,485	,485

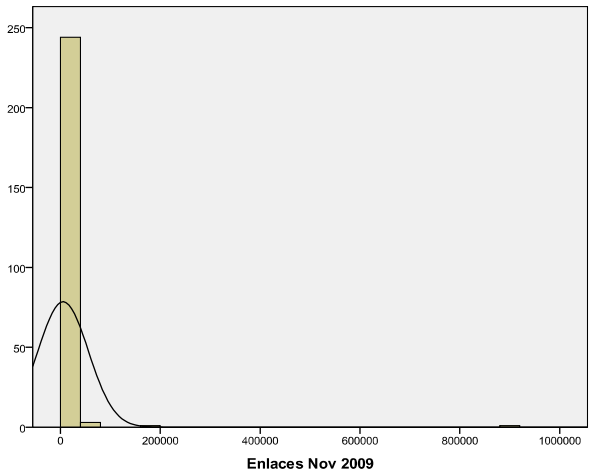
Figure 1. Histogram of the dependent variable (“inlinks”) before and after the log transformation

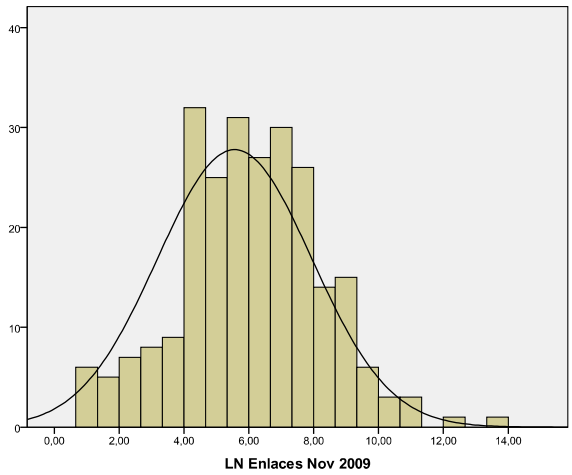
Industry: *Construction*



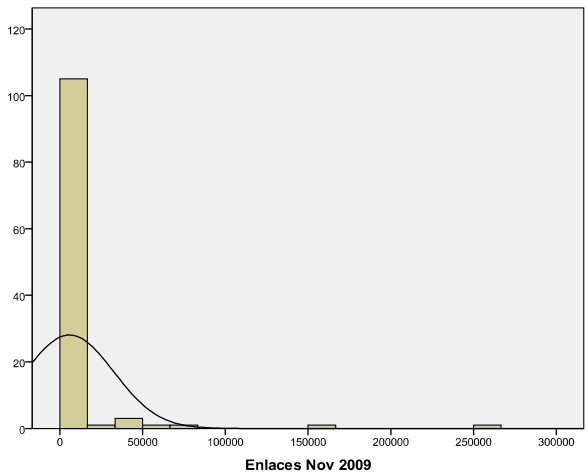


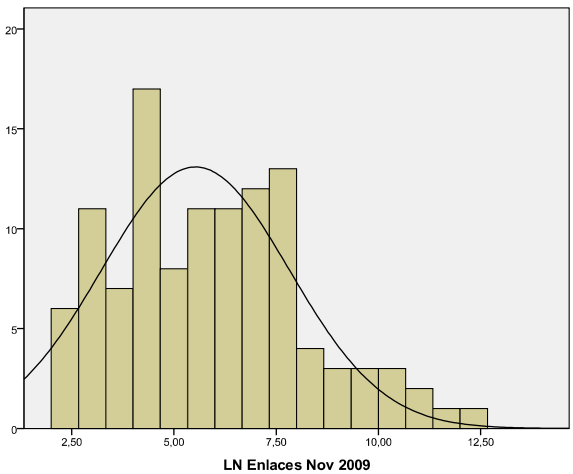
Industry: *Accommodation and Food*



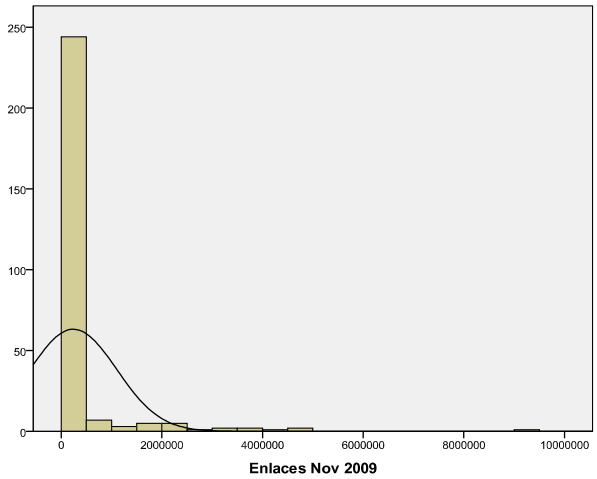


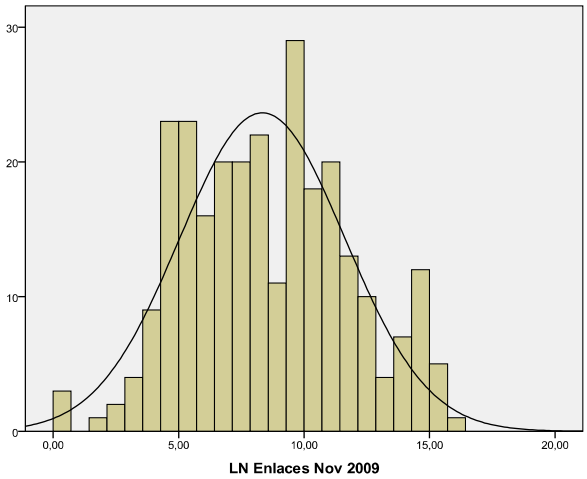
Industry: *Utilities*



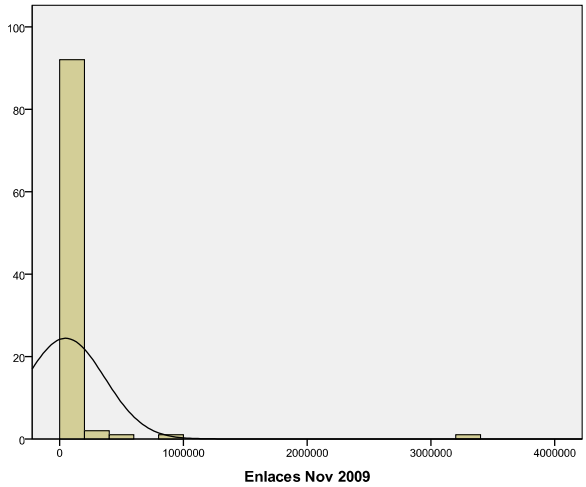


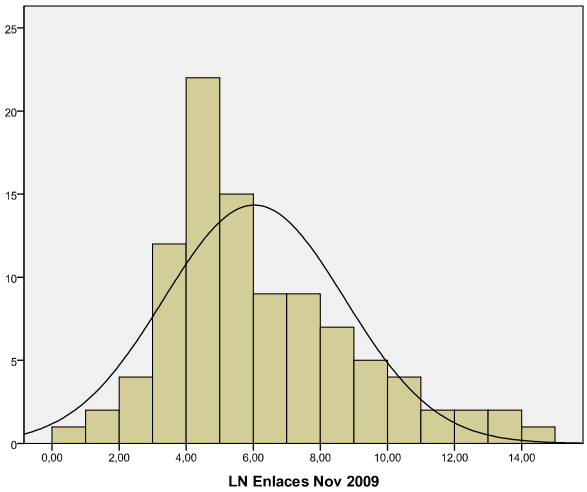
Industry: *Publishing*





Industry: *Telecommunications*





After the transformation the standardized residuals approximates a normal distribution. We analyzed some outliers (casewise diagnostics) and decided to exclude some of them due to several problems with their Web sites. The number of cases excluded were: 1 in the Telecommunications industry, 3 in Construction, 4 in Accommodation and food, 4 in Utilities and 0 in Publishing.

3.- Results and Discussion

Table 4 shows the models obtained applying the *Stepwise* method. The results were confirmed by the *Enter* method.

Table 4. Multiple regression model for the 5 industries (Stepwise method)

Industries	Constant	LN Assets	LN Intan-gible assets	LN Profit after tax	LN Turnover	Country	F	R^2
Construction	-4,494***	,582***			,274**	-1,117***	36,636***	,398
	(,974)	(,101)			(,098)	(,280)		
		,446***			,215**	-,256***		
	-4,613	5,790			2,790	-3,988		
Accommodation and Food	-1,353	,696***					73,870***	,248
	(,884)	(,081)						
		,498***						
	-1,531	8,595						
Utilities	-,874	,554***					44,311***	,305
	(,998)	(,083)						
		,552***						
	-,875	6,657						
Publishing	3,763*				1,132***	1,344**	33,561***	,209
	(1,622)				(,151)	(,402)		
					,417***	,186**		
	-2,320				7,478	3,342		
Telecommu-nications	17,126	,794***		-1,336*		-1,831***	19,187***	,401
	(8,771)	(,110)		(,612)		(,502)		
		,632***		-,184*		-,320***		
	1,953	7,231		-2,184		-3,650		

Dependent variable: LN inlinks

The cells include the following information: regression coefficient, standard error (in brackets), standardized regression coefficient and t.

***. Variable significant at the 0.001 level.

**. Variable significant at the 0.01 level.

*. Variable significant at the 0.05 level.

A summary of the final models is shown in table 5.

Table 5. Explanatory models of the five industries

Industry	Models	R^2
Construction	-4,494 + 0,582 LN Assets + 0,274 LN Turnover -1,117 Country	,398
Accomm. and Food	-1,353 + 0,696 LN Assets	,248
Utilities	-,874 + 0,554 LN Assets	,305
Publishing	3,763 + 1,132 LN Turnover + 1,344 Country	,209
Telecommu-nications	17,126 + 0,794 LN Assets -1,336 LN Profit after tax - 1,831 Country	,401

These are the main conclusions drawn from the results:

- The variable LN total assets is the most relevant variable to explain the number of inlinks received by the business Web sites. This variable is not significant only in the publishing industry. The significance level is 10,4% higher than threshold of 5%. However we consider that it is close and that this variable should consistently appear in a explanatory model of the number of inlinks received by business Web sites. If we also look at the standardized coefficients, we can observe that LN total assets is the most important variable in all the models where is significant.
- The variable LN intangibles is not significant in any industry. There could be several explanations for this: inconsistencies in the way that companies report this information, difficulties by accounting standards to recognize and measure intangible assets and lack of relationship between the intangible assets recognized in the books and the intangible assets of the company related to the Web.
- LN Turnover is included in the models for the Construction and the Publishing industries; whereas LN Profit after tax only in the telecommunications industry, but the significance level is very low as well as the standardized coefficient, 0,184.
- In general terms, LN total assets (variable of financial position) and LN Turnover (variable of performance position) are the most relevant variables. With some exceptions, we could accept hypothesis 1 and 3 and reject hypothesis 2 and 4.
- Regarding hypothesis 5, the country origin of the company is significant in 3 out of 5

models. In the construction and telecommunications industries the fact that a company is Spanish implies a lower number of inlinks. The opposite happens for the publishing industry.

- The interpretation of the coefficients is made in terms of percentages. For instance, in relation to the construction industry, an increase of 1% in the total assets implies an increase of 0,582% in the number of inlinks.
- The interpretation of the coefficients is made in terms of percentages. In instances where both the dependent variable and independent variable are log-transformed variables, the relationship is commonly referred to as elastic in econometrics. In a regression setting, we'd interpret the elasticity as the percent change in y (the dependent variable), while x (the independent variable) increases by one percent. For instance, in relation to the construction industry model, a one percent increase in the total assets would yield a 0,582% increase in the number of inlinks.

Finally we have checked the assumptions included in the table 1. The diagrams are included in the Appendix 1 by industries. The histograms of the standardized residuals and the P-P plots of normally distributed residuals show that the normality condition can be accepted. The plots of standardized residuals against standardized predicted values indicates that there is no significant problem of heterocedasticity. Finally the partial plots show that in general terms there is a linear relationship between dependent and independent variables. Multicollinearity tests (Tolerance and VIF) are shown in table 6. There are no multicollinearity problems.

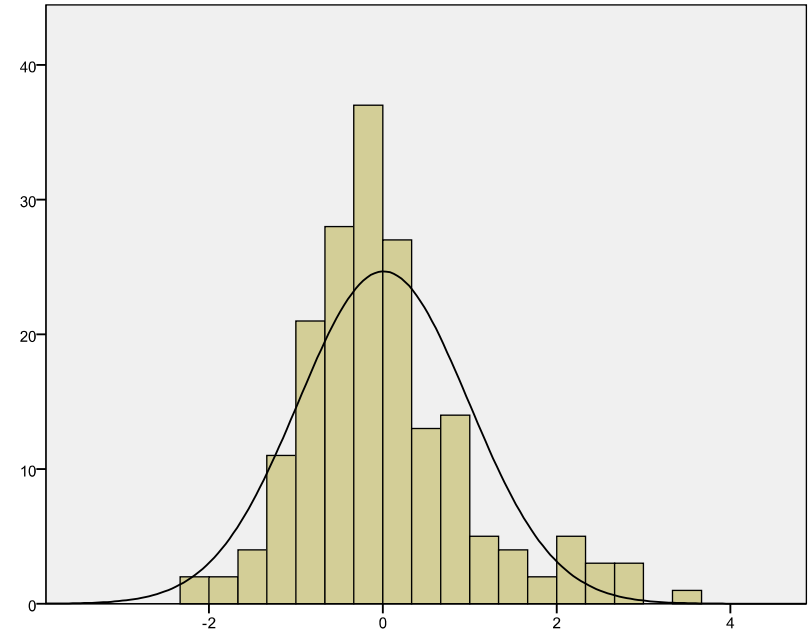
Table 6. Multicollinearity tests

Industry	Variables	Tolerance	VIF
Construction	LN Assets	,612	1,635
	Country	,878	1,139
	LN Turnover	,613	1,632
Accommodation and Food	LN Assets	-	-
Utilities	LN Assets	-	-
Publishing	LN Turnover	1,000	1,000
	Country	1,000	1,000
Telecommunications	LN Assets	,913	1,095
	Country	,904	1,107
	LN Profit after tax	,987	1,013

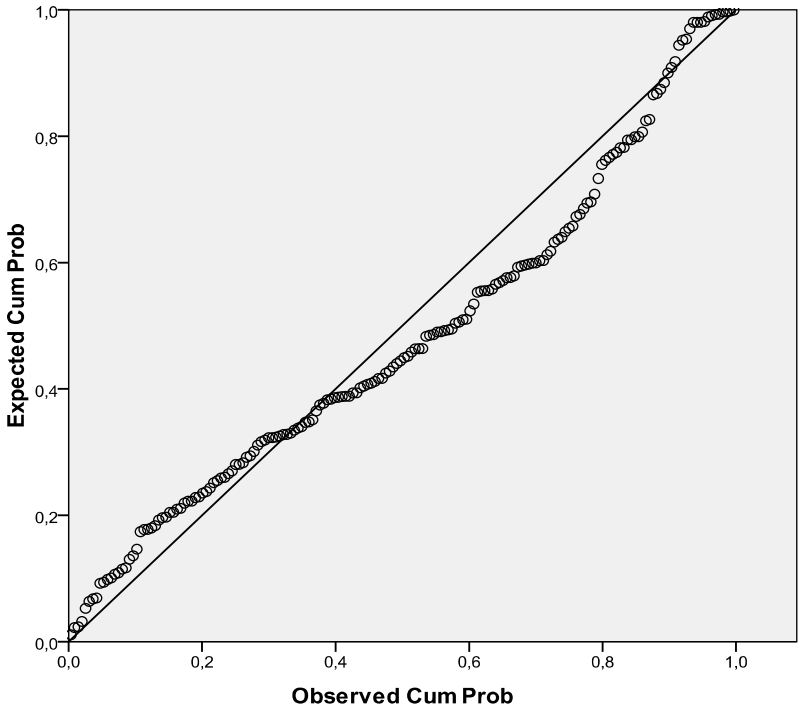
Appendix

Industry: *Construction*

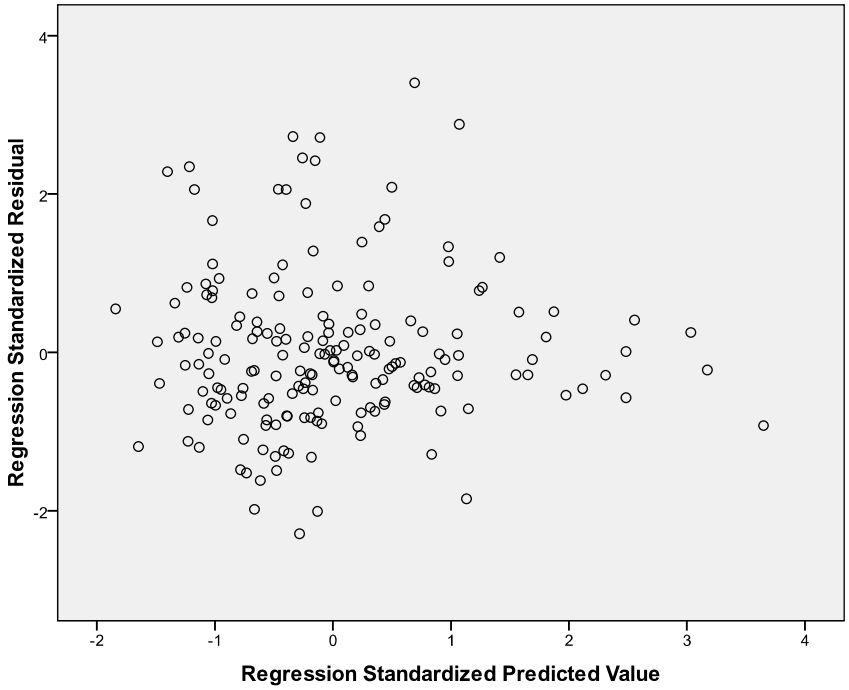
Histogram of standardized residuals



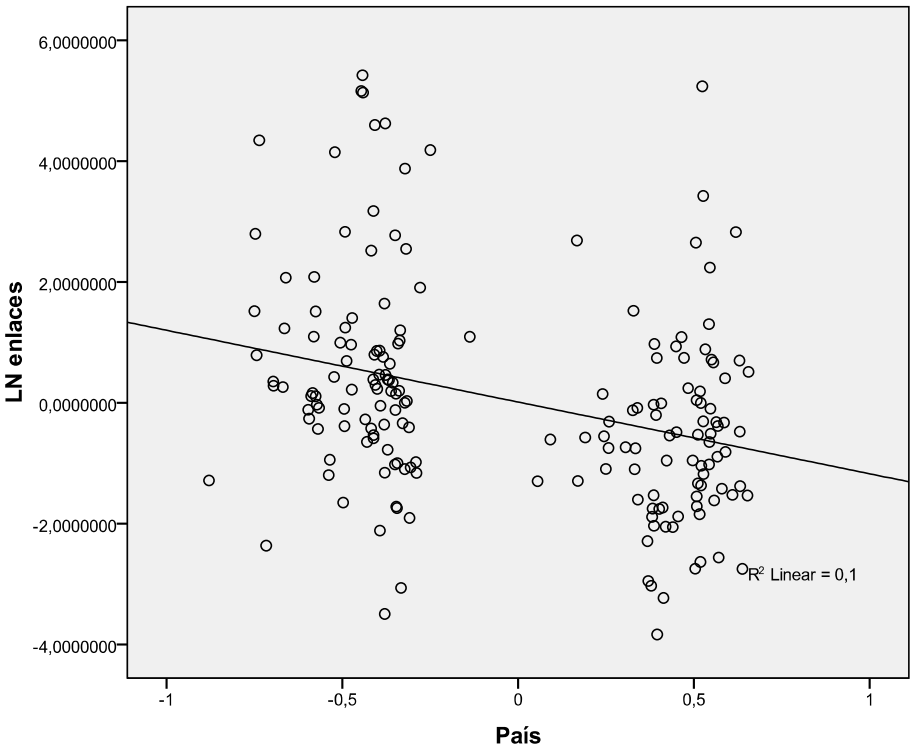
P-P plots of normally distributed residuals

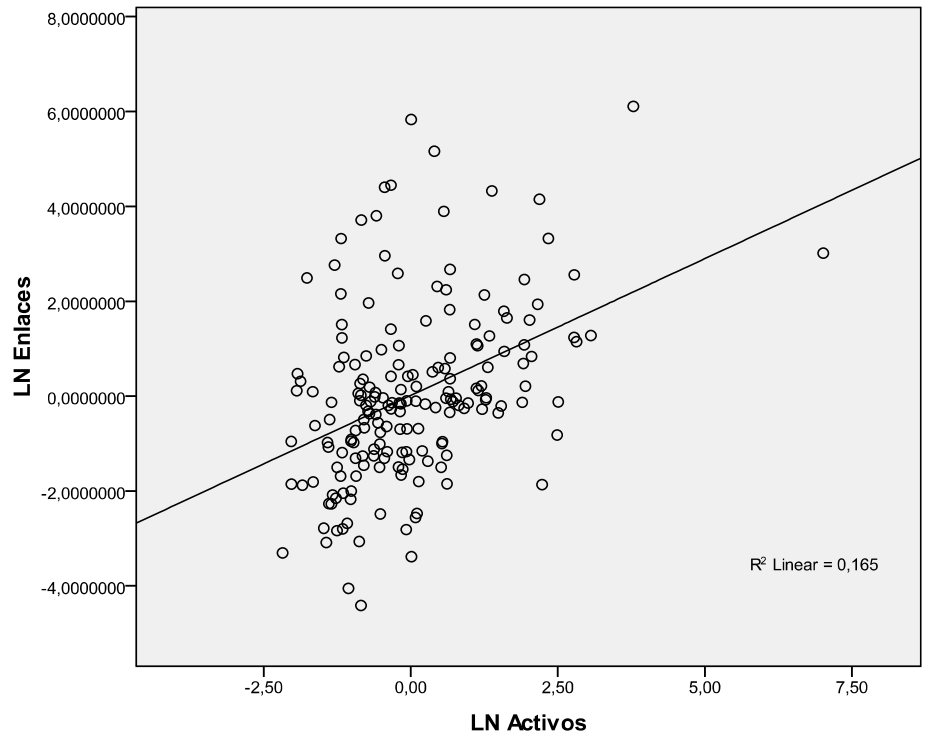


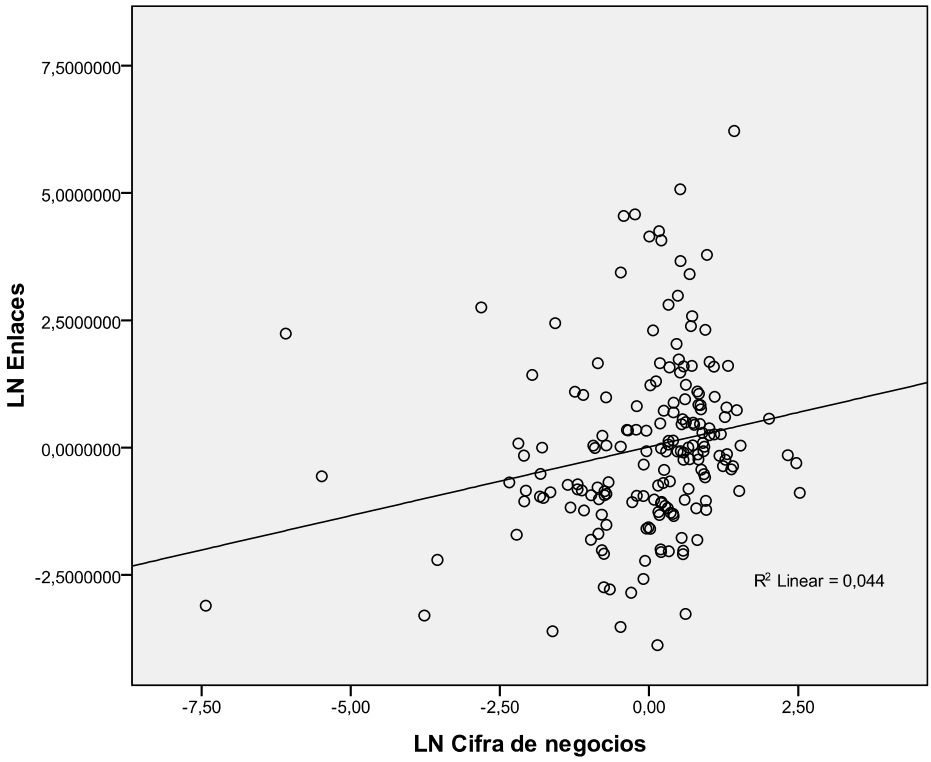
Plot of standardized residuals against standardized predicted values



Partial plots

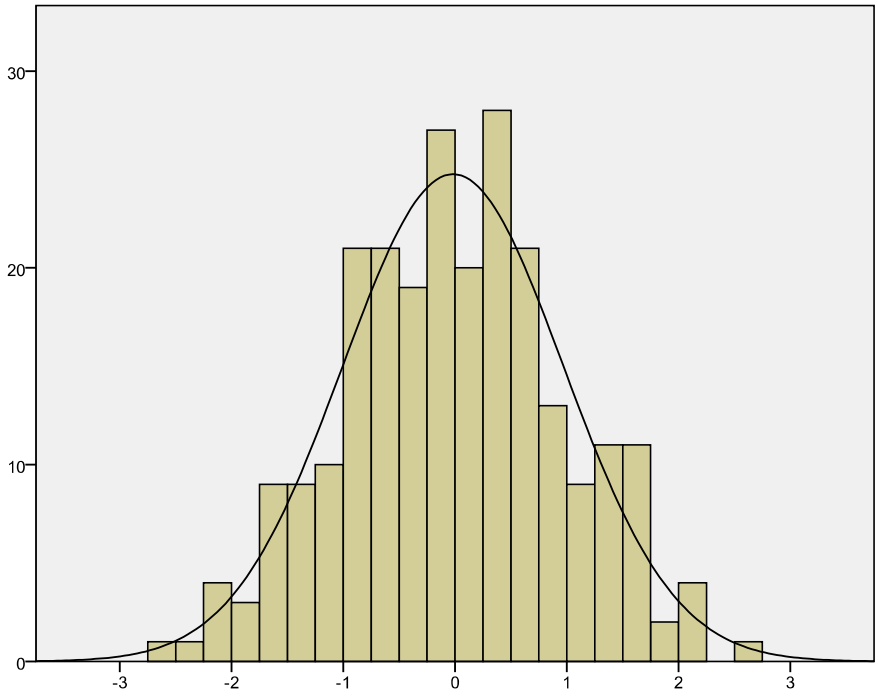



8,0000000
6,0000000
4,0000000
2,0000000
0,0000000
-2,0000000
-4,0000000
LN Enlaces
-2,50
0,00
2,50
5,00
7,50
LN Activos
R² Linear = 0,165

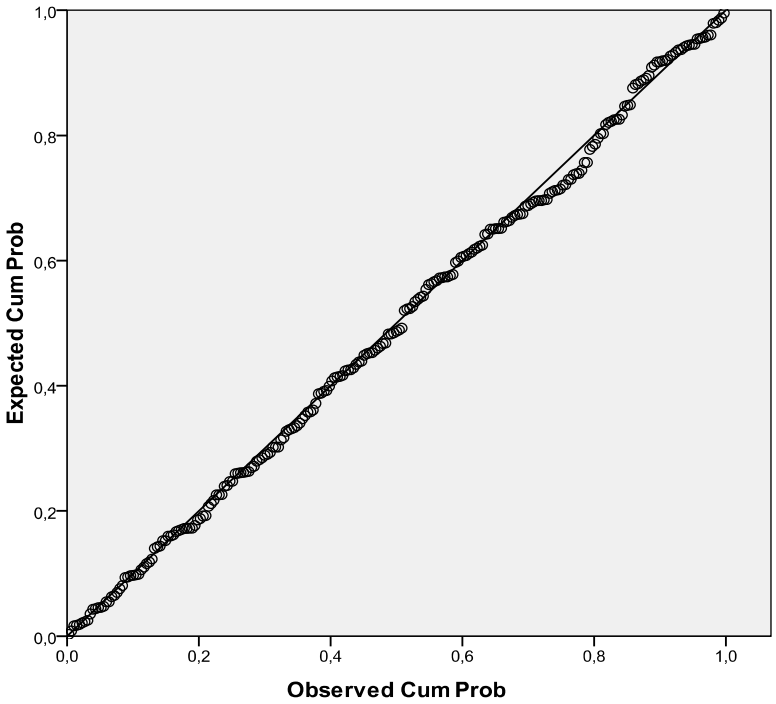

7,5000000
5,0000000
2,5000000
0,0000000
-2,5000000
LN Enlaces
-7,50
-5,00
-2,50
0,00
2,50
LN Cifra de negocios
R² Linear = 0,044

Industry: *Accommodation and Food*

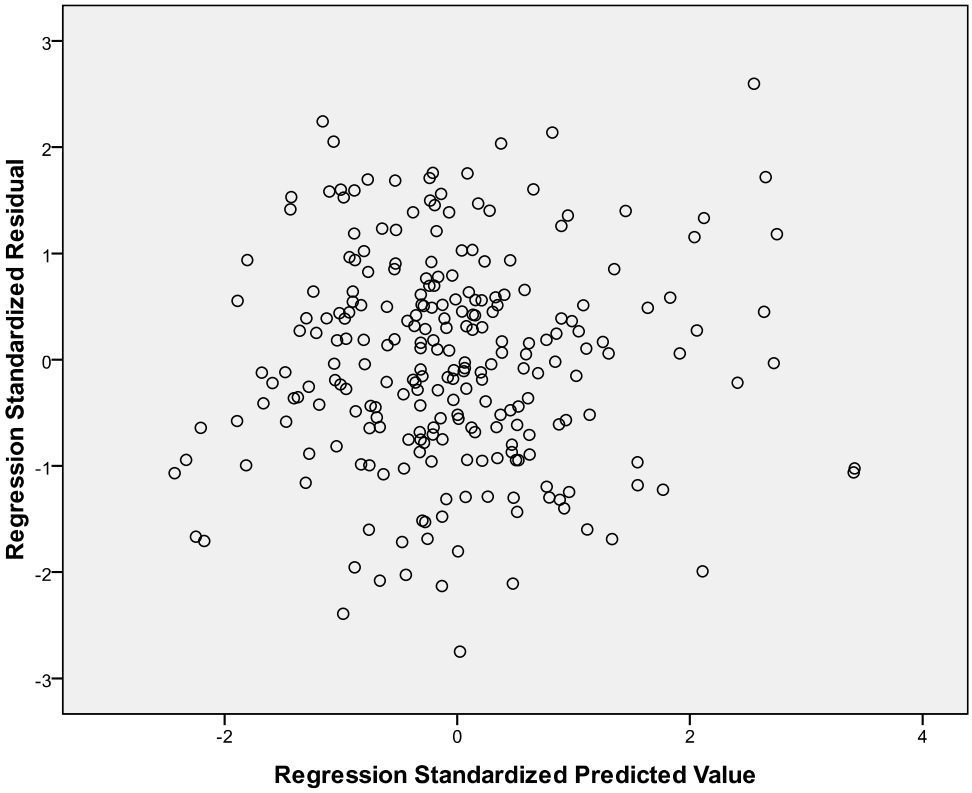
Histogram of standardized residuals



P-P plots of normally distributed residuals

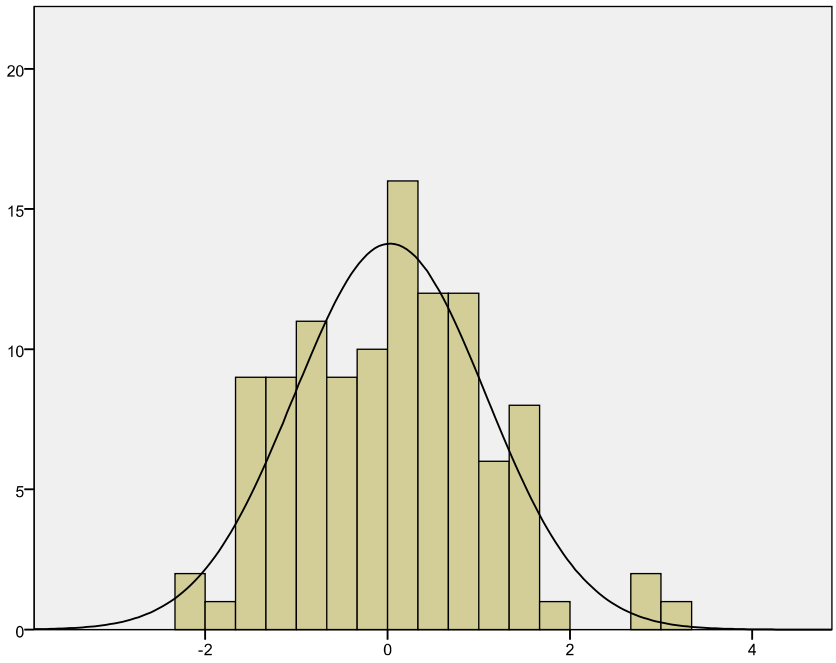


Plot of standardized residuals against standardized predicted values

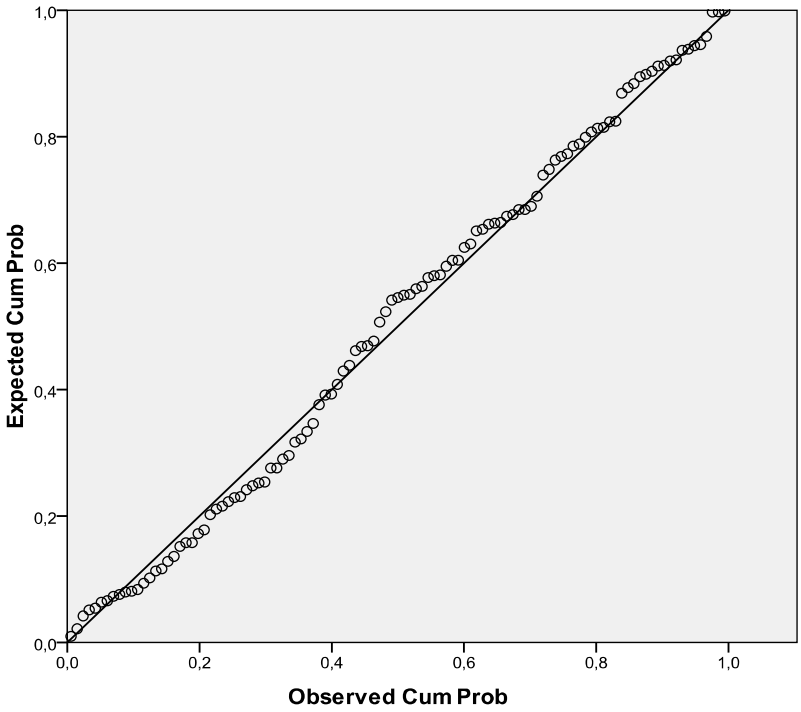


Industry: *Utilities*

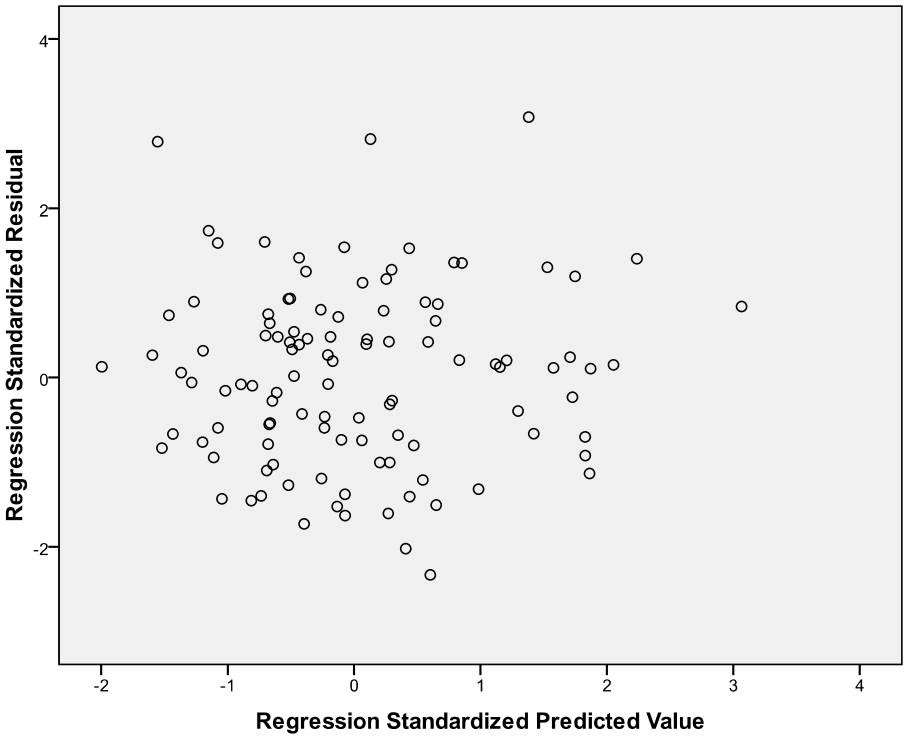
Histogram of standardized residuals



P-P plots of normally distributed residuals

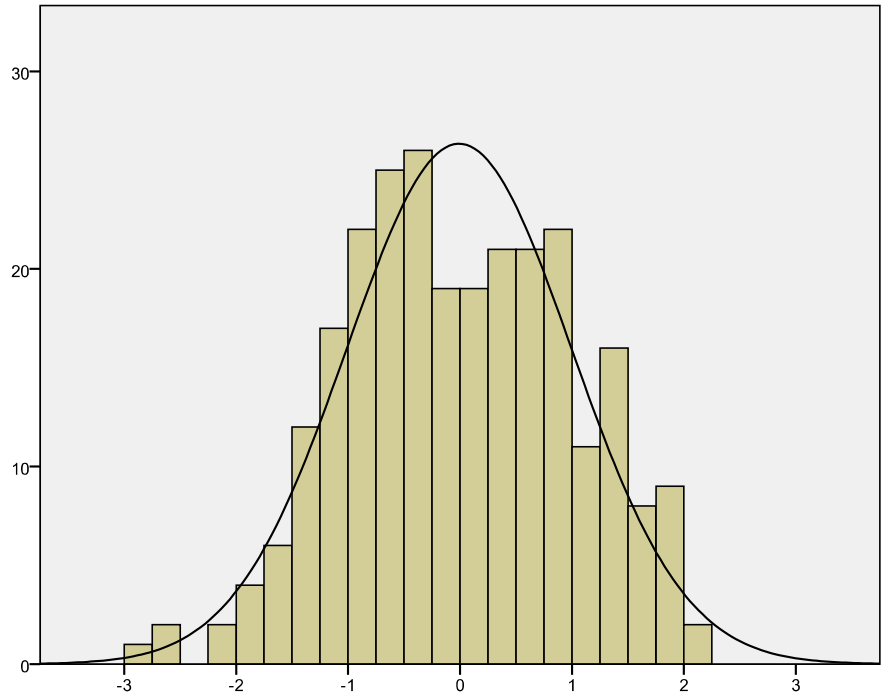


Plot of standardized residuals against standardized predicted values

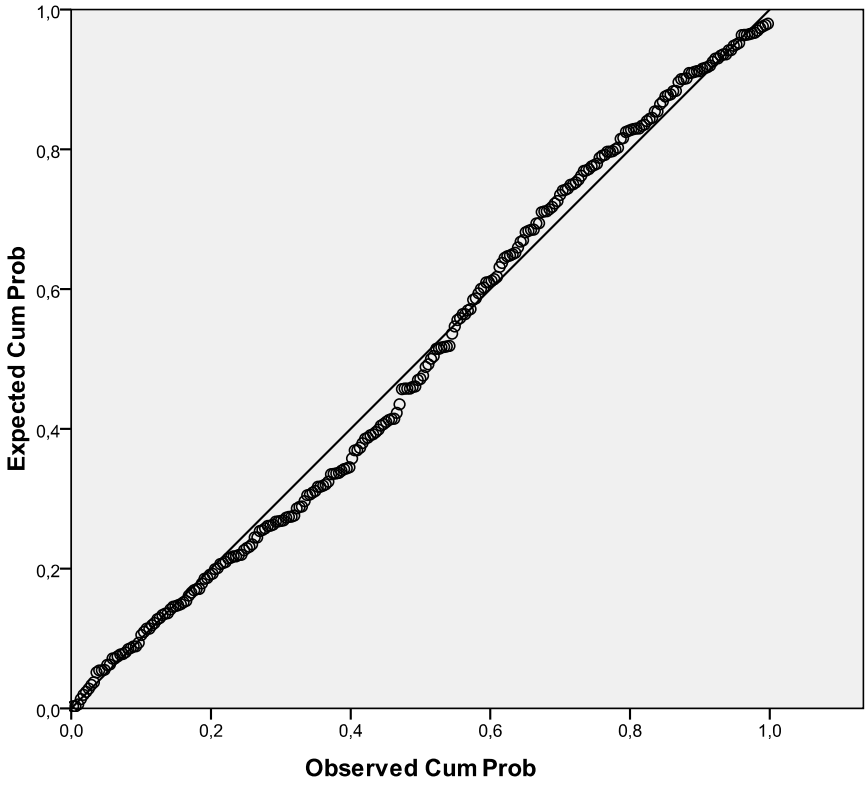


Industry: *Publishing*

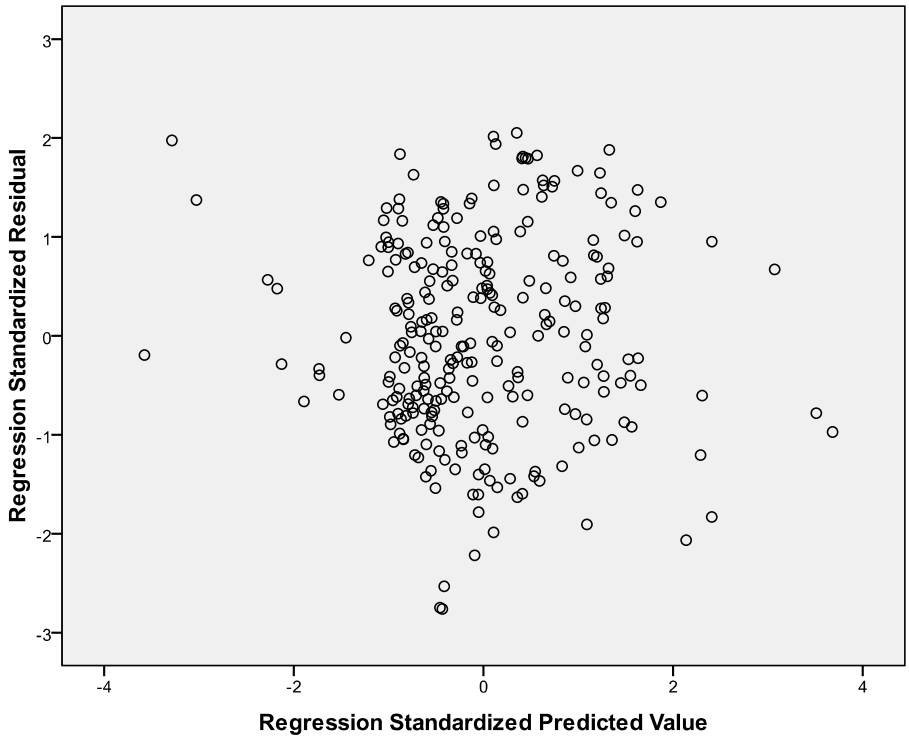
Histogram of standardized residuals



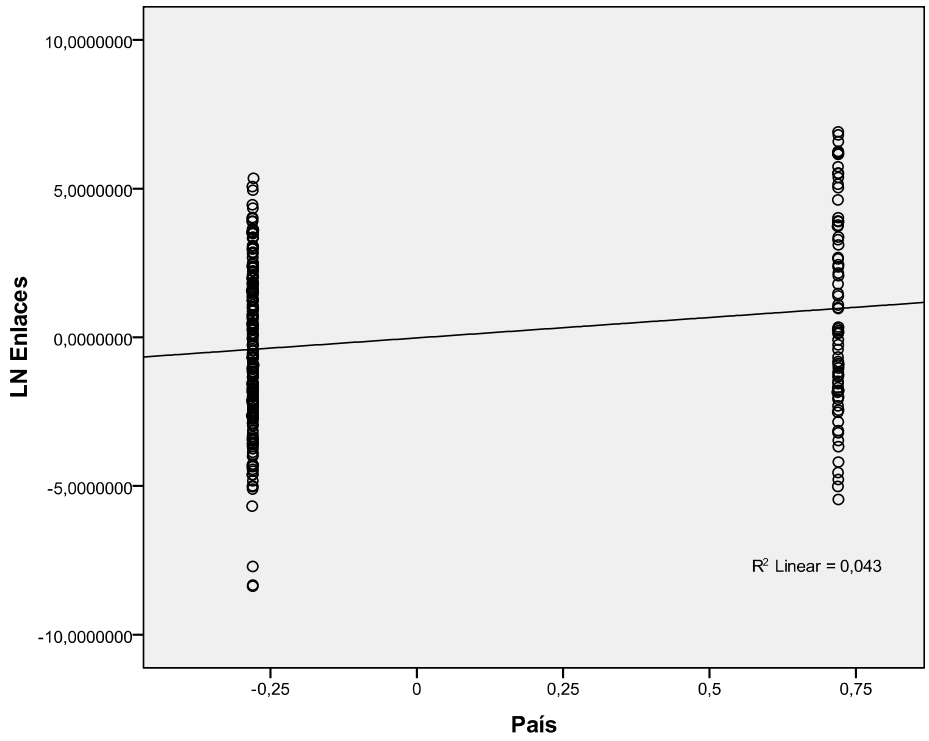
P-P plots of normally distributed residuals

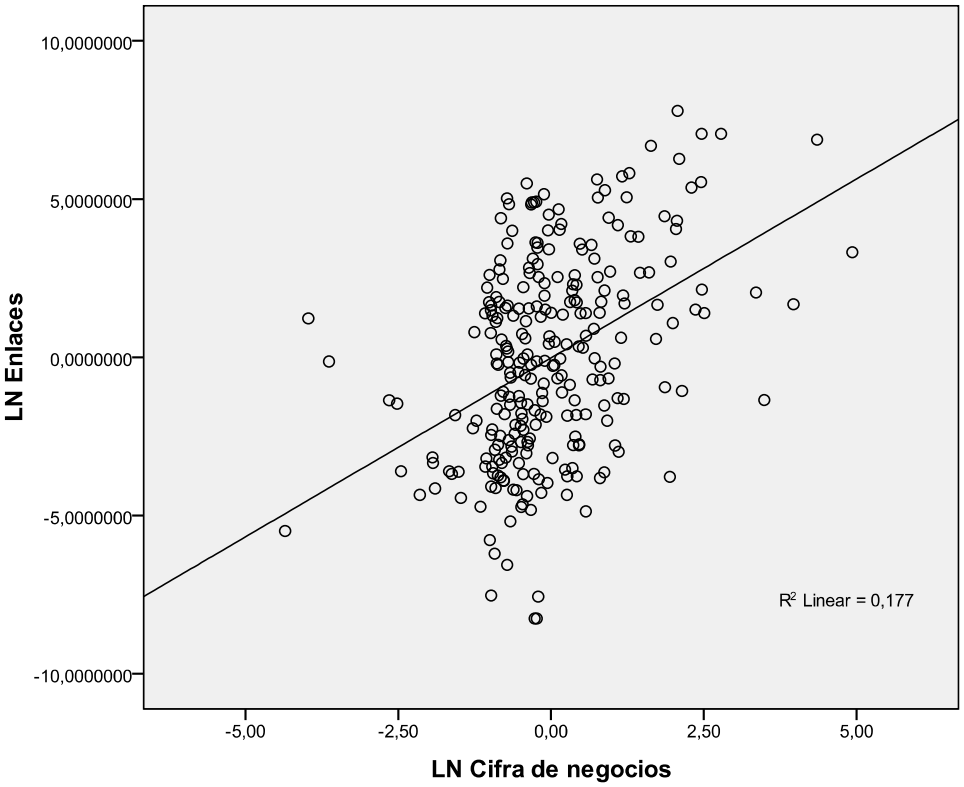


Plot of standardized residuals against standardized predicted values



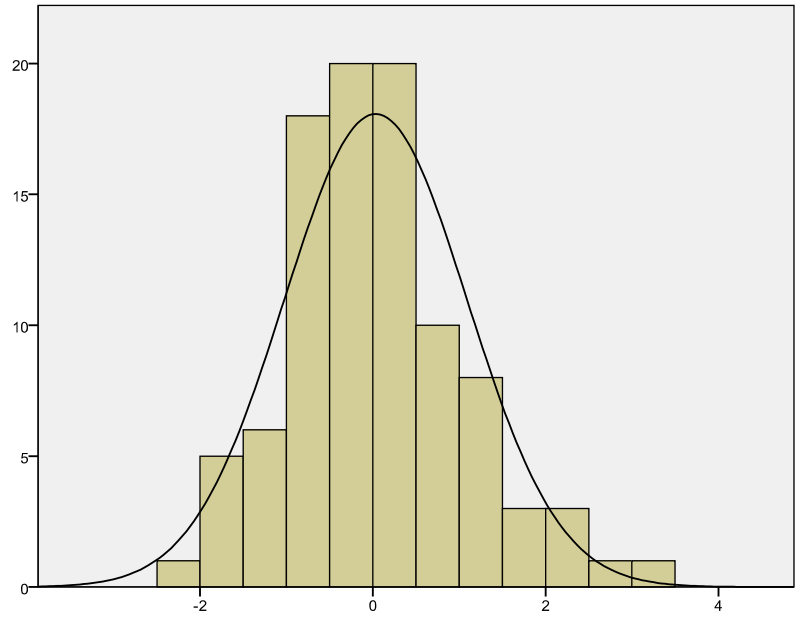
Partial plots



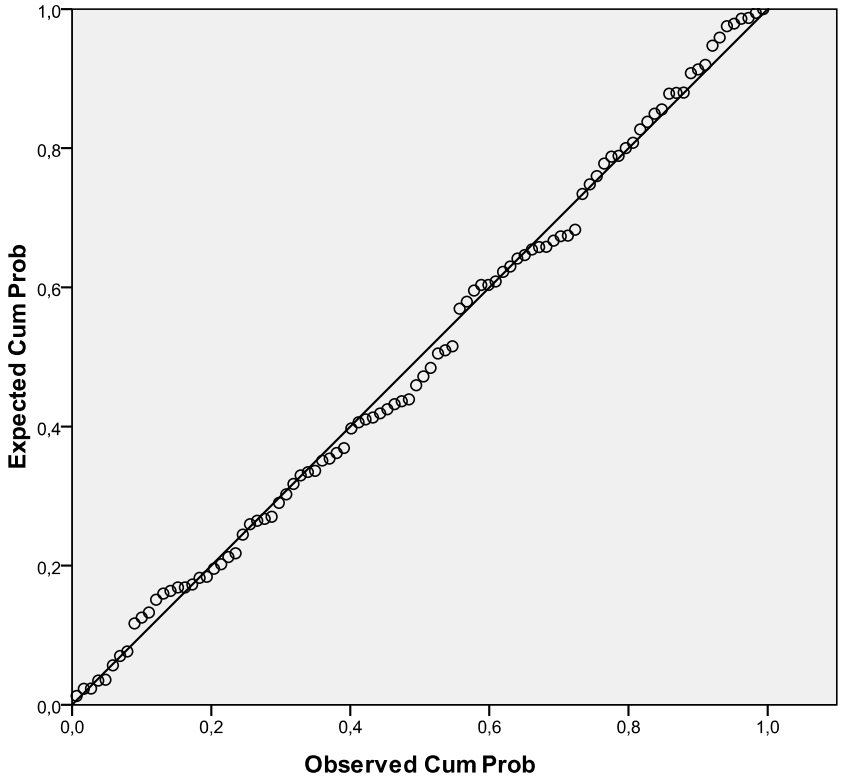

10,0000000
5,0000000
0,0000000
-5,0000000
-10,0000000
LN Enlaces
-5,00
-2,50
0,00
2,50
5,00
LN Cifra de negocios
R² Linear = 0,177

Industry: *Telecommunications*

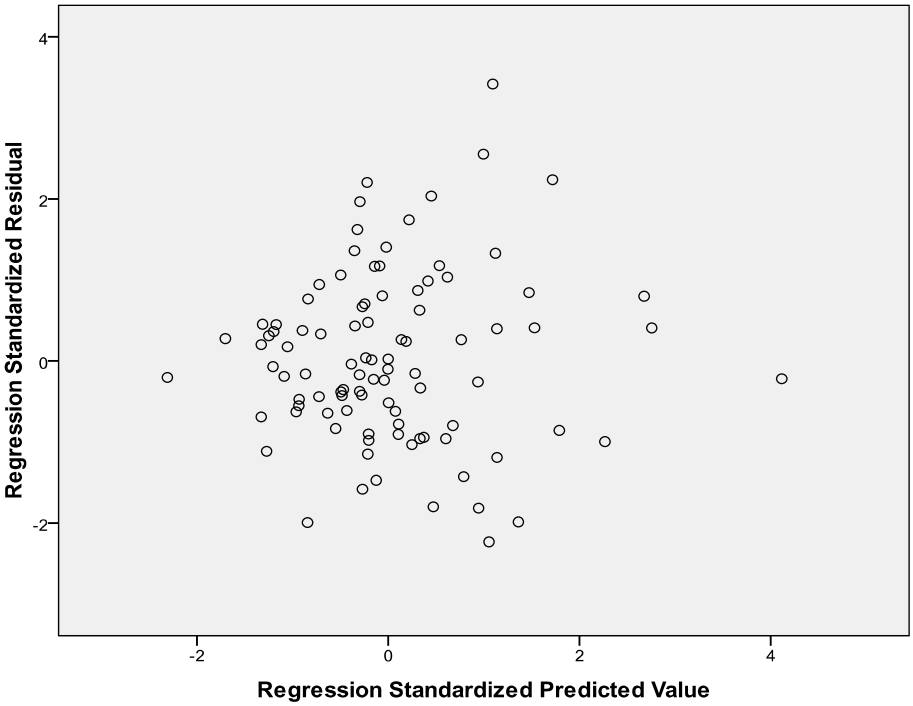
Histogram of standardized residuals



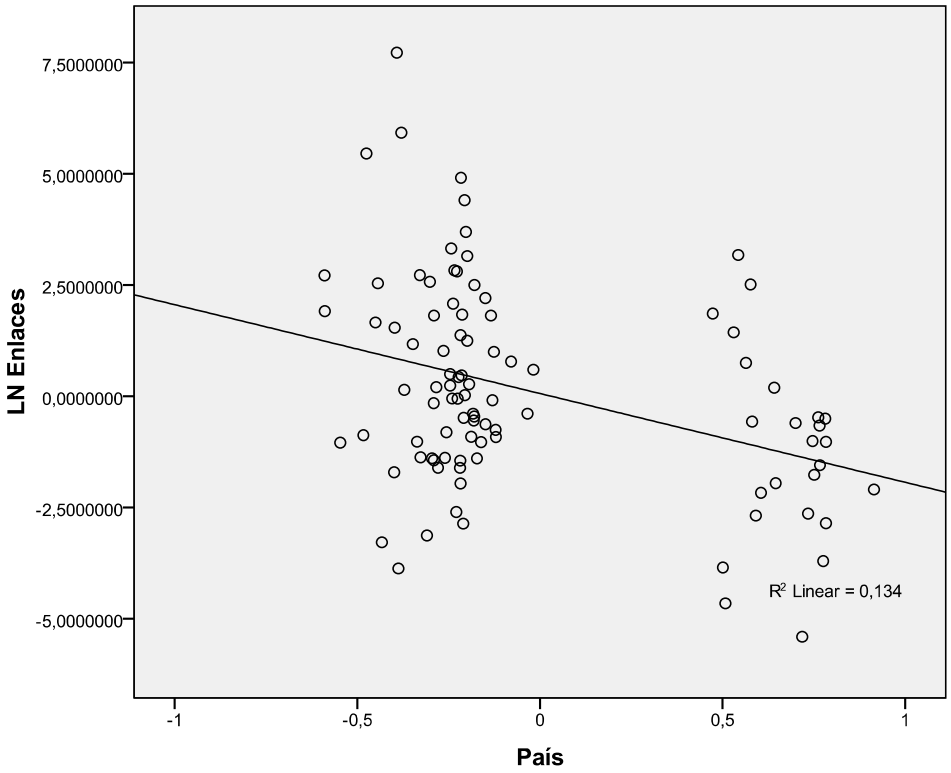
P-P plots of normally distributed residuals

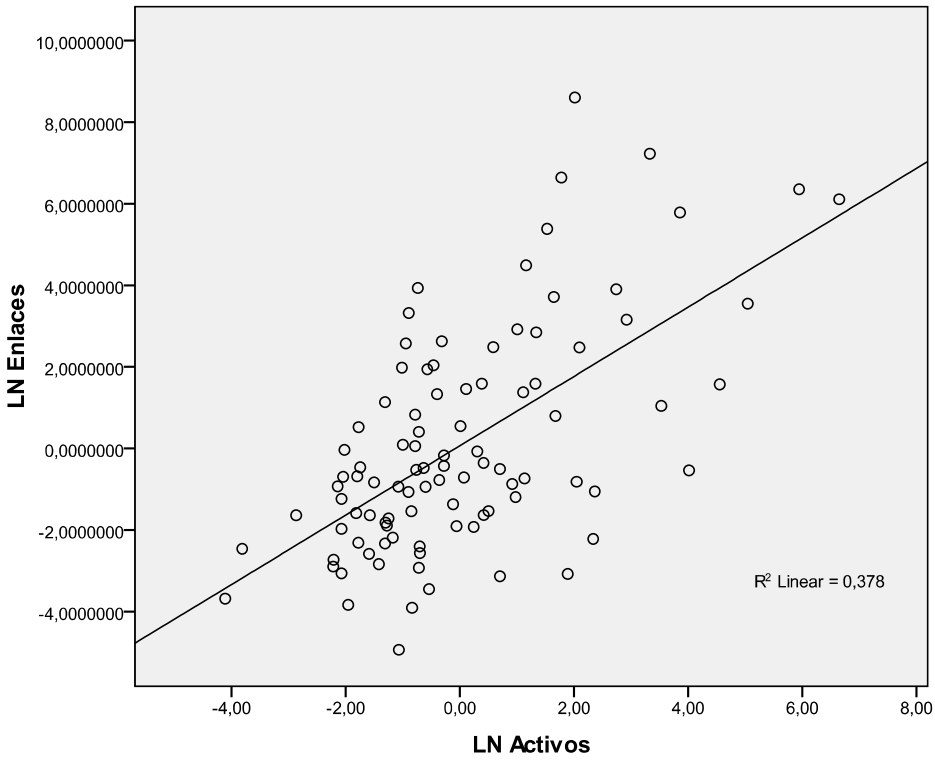


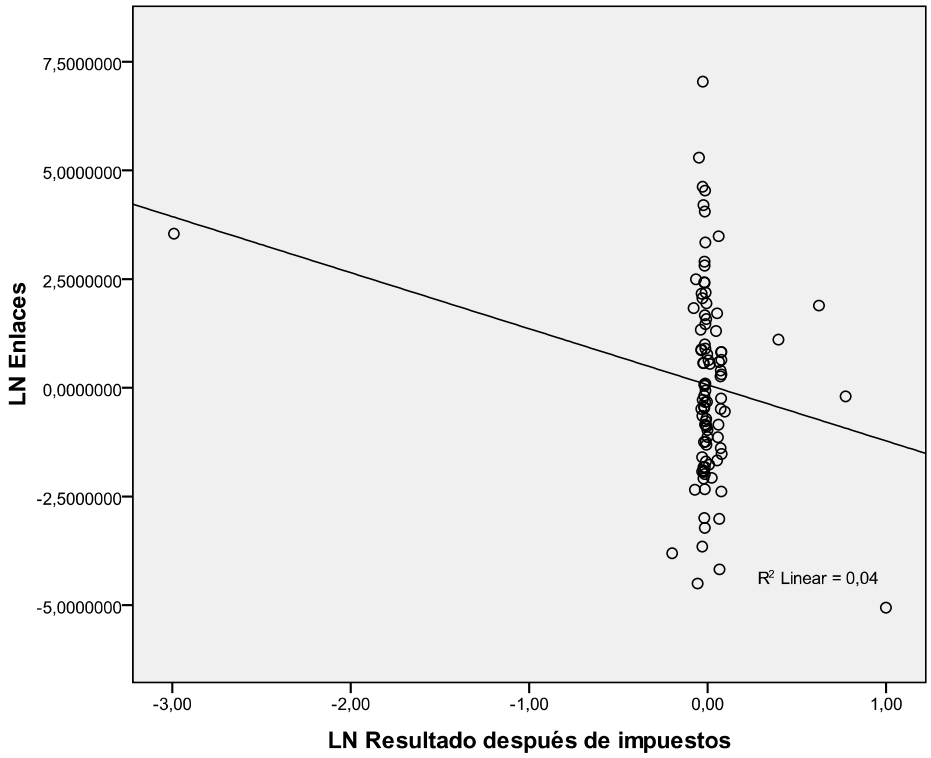
Plot of standardized residuals against standardized predicted values



Partial plots




LN Enlaces
10,0000000
8,0000000
6,0000000
4,0000000
2,0000000
0,0000000
-2,0000000
-4,0000000
-4,00
-2,00
0,00
2,00
4,00
6,00
8,00
LN Activos
R² Linear = 0,378


LN Enlaces
7,5000000
5,0000000
2,5000000
0,0000000
-2,5000000
-5,0000000
-3,00
-2,00
-1,00
0,00
1,00
LN Resultado después de impuestos
R² Linear = 0,04

3.2 Co-link analysis

3.2.1 Co-link analysis of heterogeneous companies belonging to some of the main stock exchange indexes in the world

The reference of the paper included in this section is:

ROMERO-FRÍAS, E. & VAUGHAN, L. (2010) “Patterns of Web Linking to Heterogeneous Groups of Companies: The Case of Stock Exchange Indexes”. To appear in *Aslib Proceedings*.

Patterns of Web Linking to Heterogeneous Groups of Companies: The Case of Stock Exchange Indexes

Abstract

Purpose – To extend co-link analysis to Websites of heterogeneous companies belonging to different industries and countries. To cluster companies by industries and compare results from different countries.

Design/methodology/approach – Websites of 255 companies that belong to five stock exchange indexes were included in the study. Data on co-links pointing to these Websites were gathered using Yahoo!. Co-link data were analyzed using multidimensional scaling (MDS) to generate MDS maps that would position companies based on their co-link counts.

Findings –Comparisons of results across different countries and economies showed the following overall pattern: companies whose businesses are information based tend for form well defined clusters while companies operating on a more traditional business model tend not to form clear groups. A comparison between EU zone and the U.S. suggests that the EU economy is not well integrated yet.

Practical implications – Findings from the study suggest the possibility of using co-link analysis to distinguish the information based industries from traditional industries.

Originality/value –Extended co-link analysis from a single industry to heterogeneous industries with global and complex business phenomena.

Keywords Competitive intelligence, Web data mining, co-link analysis, Webometrics

Paper type Research paper

1. Introduction

Thirty years ago Michael Porter (1980, 1985) developed the idea that the creation and maintenance of competitive advantages are key factors of business success. Despite of the evolution of business theories over the last three decades, competitive strategy remains as a fundamental element in the management of a company. Competitive Intelligence (CI) is the result of the general acceptance of this important managerial thinking. Today CI plays a key role in a company's strategic effort of achieving and sustaining its competitive advantages (Heppes and du Toit, 2009). According to Kahaner (1996), CI consists of a systematic plan to obtain and analyze information about competitors and general trends in the industry. The abundance of information on developed economies, especially with the development of digital technologies and the Internet, has created new opportunities and challenges for companies so that they need to monitor changes around them in order to compete in better conditions. In the last decade, Web hyperlink analysis has been used for CI purposes with promising results (Reid, 2003; Tan *et al*., 2002).

Inlink and co-inlink are basic Webometric concepts to understand the nature of Web hyperlinks. According to Björneborn and Ingwersen (2004), an inlink, also called back link, is a link pointing to a Webpage, e.g. page X has an inlink coming from page Z. If page X and page Y both have inlinks from page Z, then page X and page Y are co-inlinked. Co-inlink analysis is referred simply as co-link analysis later in this paper.

Previous research has established the relationship between the number of links pointing to a business website and the company's financial performance, e.g. revenue and profit (Vaughan and Wu, 2004). Moreover, direct links pointing to websites have proved to be a relevant measure of similarity (Thelwall and Wilkinson, 2004). However, business competitors link each other very rarely in order to avoid diverting Web traffic to rival companies (Shaw, 2001; Vaughan, Gao and Kipp, 2006). This fact leads to the use of co-links rather than inlinks to study competitive positions. As Vaughan (2006) points out, co-link data are more robust than inlink data as the former are less easily manipulated. Content analysis has been carried out to study the motivations for co-link creation (Vaughan, Kipp and Gao, 2007). Co-link analysis has also been demonstrated to be a useful tool to reveal the cognitive or intellectual structure of a particular field of study (Zuccala, 2006), analogously to co-citation analysis in bibliometrics (Small, 1973).

Web co-link analysis for business information started with a single industry (Vaughan and You, 2006) or on a detailed picture of the competitive landscape of a sector within an industry (Vaughan and You, 2008; Vaughan and You, 2009). The methodology developed in these papers has also been tested and verified in other countries and other industries, e.g. China's chemical industry and electronics industry (Vaughan, Tang and Du, 2009). Romero-Frías and Vaughan (2009) extended the use of co-link analysis into the banking industry in the US in order to test the feasibility of combining page content with co-link data to monitor financial crisis. Parallel to these studies on

commerical Websites, research on heterogenous Websites has been carried out to test the triple helix theory on the Web (Stuart and Thelwall, 2006; García-Santiago and de Moya-Anegón, 2009).

This paper reports a study that extended co-link analysis to Websites of heterogeneous companies belonging to five stock exchange indexes. These companies are the biggest in their respective economies and therefore have a significant presence on the Web and are likely to receive attention from users, companies, government and other economic agents. When studying a single industry, co-link analysis allowed us to visualize business competition within the industry. When we applied co-link analysis to companies belonging to different industries in this study, we may see not only competition within an industry but also alliances or other relationships between different industries. The scope of this study is much broader than those carried out before, which gives us an opportunity to explore the issues that were not examined in previous co-link studies. We compared results from the five different stock exchange indexes that represent different economic, geographical and cultural backgrounds. We analyzed the Eurostoxx 50 from two different perspectives of country and industry. This allowed us to observe the degree of economic integration in the Euro-market. We also tried to observe if there is any clear pattern, common to all the stock exchanges, about the type of business activities that attract more Web links and other activities that attract fewer Web hyperlinks due to their particular business features. We observed that there is one main variable that could determine a generalised pattern between economic activities, this is, the degree in which the activity is information centered. This applies mainly to IT, media, and financial companies, among others. These industries are in an ongoing process of business model transformation, e.g. the deep crisis in the media sector due to digitalisation of information. The paper is exploratory in nature with an intent to open new lines of research on the application of Webometric methodology to business studies.

2. Methodology

2.1. Selection of companies to study

All companies in the following stock exchange indexes are candidates of the study: Dow Jones Industrial (30 US companies), the Dow Jones Euro Stoxx 50 (Euro zone companies), CAC 40 (Paris), FTSE 100 (London), and Ibex 35 (Madrid). The complete list of all these 255 companies is omitted due to a space consideration; however, a selection of companies is included in Appendixes 1 to 5 to facilitate the understanding of the findings (Figures 1 to 7). Table 1 provides an overview of the indexes in the study: country, number of companies, composition retrieval date, URL and the appendix in which the companies are listed. Readers can refer to the URLs in the Table for the companies in each of the indexes, however changes in the composition are expected over the

time. In order to facilitate the identification of companies that are omitted in the appendixes, the company labels in the MDS maps are the tickers used in the stock exchanges for each business.

Table 1. Information about the stock exchange indexes in the study

Index	Country / Region	Number of companies	Composition retrieval date	URL	Appendix
Dow Jones Industrial	U.S.A.	30	17/02/2009	http://www.djaverages.com/	1
FTSE 100	U.K.	100	24/02/2009	http://uk.finance.yahoo.com/q/cp?s=^FTSE	2
Eurostoxx 50	12 Eurozone countries	50	17/02/2009	http://www.stoxx.com/indices/components.html?symbol=SX5E	3
CAC 40	France	40	19/02/2009	http://www.euronext.com/trader/indicescomposition/composition-4411-EN-FR0003500008.html?selectedMep=1	4
IBEX 35	Spain	35	19/02/2009	http://www.bolsamadrid.es/ing/mercados/acciones/accind1_1.htm	5

The Website address of each of these companies was collected from the stock exchange Website and then manually checked to ensure its correctness. The vast majority of companies in the study have only one URL for their Websites. For the few that have alternative URLs in the form of alias or redirect, we checked each URL to find out which one has more inlinks and used that one for collecting inlink data. We considered including both URLs in data collection. However, Yahoo!, the search engine used for data collection, cannot handle the complex query syntax for collecting co-link data using two URLs.

Industry classification in each stock exchange differs. For example, the Industry Classification Benchmark (ICB) is used globally (though not universally) to divide the market into increasingly specific categories. Currently it is used by Dow Jones, FTSE and several other markets around the globe. However, other markets, e.g. the Spanish one represented by IBEX 35, use its own industrial classification. This heterogeneity, together with the scarce number of companies for some industries in the indexes, obliges to merge companies into general groups such as IT, Media or Financial. We understand that this classification used for practical purposes in the study is

subject to discussion. Therefore, brief notes about the native industrial classification for each company is provided in Appendixes 1 to 5.

2.2. Collection of Web link data

Of the three major search engines, Google, Yahoo! and MSN Live Search, only Yahoo! could be used for data collection for the study. Google's inlink search only returns a sample of all inlinks that the Google database records (Google, 2009). Another problem is that Google cannot filter out internal inlinks (inlinks originated from within the Website itself such as "back to home" type of links) as the query term 'link" cannot be combined with any other query terms (Google, 2006). In other words, it cannot report the external inlink counts that the study needed. MSN Live Search used to have inlink search functions but the service was turned off around March 2007 (Live Search, 2007). At the time of data collection, winter 2009, Yahoo! is the only option for collecting inlink data as required by the study.

Because search engines of different countries may have databases that favour Websites of the host countries (Vaughan and Thelwall, 2004), we considered using the Spanish version of Yahoo! to retrieve inlinks to the companies in the Spanish stock exchange and the French version of Yahoo! for French companies. However, tests of these versions of search engine showed that they returned the same inlink search results as that from the global version of Yahoo!. Unlike the Chinese version of Yahoo! which has a database that is different from the global version of Yahoo!, the Spanish and French version of Yahoo! just had a different interface but the same underlying database. So the global version of Yahoo! (www.yahoo.com) was used for all data collection.

Yahoo! has two inlink search query terms, link and linkdomain. The "link" query term finds links to a particular page (e.g. link:http://www.abc.com finds links to the homepage of www.abc.com) while the linkdomain query term retrieves all links that point to all pages of a particular Website or domain including the homepage. We used the linkdomain query term for data collection because all links, not just links to homepage are of relevance to the study. The query syntax for the data collection is illustrated in Table 2 using the example URLs of www.abc.com and www.xyz.com. We truncated the www portion of the URLs in the queries to capture all links to all subdomains such as mail.abc.com. The "-site:abc.com" part of the query is to filter out internal links coming from within the domain of abc.com itself.

Table 2. Illustration of Yahoo! Queries for Data Collection

Types of links searched for	Query
Inlinks to www.abc.com	linkdomain:abc.com –site:abc.com
Co-links between www.abc.com and www.xyz.com	(linkdomain:abc.com –site:abac.com) (linkdomain:xyz.com –site:xyz.com)

Since co-links involve a pair of Websites, the co-link data are collected in the form of a matrix with row x and column y of the matrix representing the number of co-links between URL x and URL y. We collected co-link data by stock exchanges, i.e. companies in the same stock exchange are in the same co-link matrix. Thus there are five co-link matrices for the five stock exchanges.

2.3. Methods of data analysis

Each co-link matrix was analyzed using multidimensional scaling (MDS) to generate a MDS map. MDS uses a heuristic method to place companies with higher co-link counts closer in the resulting MDS map. Since similar or related companies are more likely to receive co-links (two unrelated companies such as a food company and an IT company has a very small chance of being co-linked), the number of co-links between a pair of companies is a measure of their relatedness. Therefore, similar or related companies will be placed closer in the MDS map. The positions of the companies in the map only reveal their relative relationship to each other, e.g. if we modify the list of companies in the analysis, the positions would also change. Therefore, it is important to highlight that there is no absolute meaning in the position of a business in the map, but only if we take into consideration its position in relation to other companies. We hoped to use MDS the map to see clusters of companies and to reveal their market positions. We also hoped to examine of the effectiveness of our methods by comparing the MDS maps from the five stock exchanges.

The raw co-link count collected from Yahoo! needed to be normalized to obtain a relative measure of the relatedness of companies. This is necessary because a co-link count of ten is large if the two companies in question each had very few inlinks, e.g. 15. On the other hand, if each company received thousands of ininks, then the 10 co-links that they shared represented a very small portion. The normalization was done by applying the Jaccard index in the following way.

Normalized co-link count $=n(A\cap B)/n(A\cup B)$

Where A is the set of the Web pages that linked to Website X

B is the set of the Web pages that linked to Website Y

$n(A \cap B)$ is the number of pages that linked to both Website X and Website Y, i.e. the raw co-link count

$n(A \cup B)$ is the number of pages that linked to either Website X or Website Y.

The normalized co-link matrices were feed into SPSS for MDS analysis. The MDS map for the FTSE 100 was too crowded to see clusters of companies clearly for a meaningful interpretation. So we reduced the number of companies from 100 to 80 and then to 35. We ranked the companies by the number of inlinks and selected the top 80 and 35 companies for co-link MDS analysis. Both the top 80 and top 35 maps are included (Figure 3 and 4) as they both are useful to understand the overall patterns discovered. One company in Euro 50, Generali, had no co-links with most other companies so it was omitted from the co-link analysis. The stress values of MDS analysis are 0.03 for Dow Jones, 0.07 for Euro 50 (49 companies), 0.06 for CAC 40, and 0.05 for Ibex 35, 0.05 for FTSE top 80 and 0.06 for top 35. These stress values are all fairly low, which suggests a good fit between the data and the MDS map positions.

3. Results

The five stock exchange indexes are representative of different economies and differ in industry composition and in social, cultural and political backgrounds. In this section we present a sketch of each stock and then compare them to obtain general findings and patterns. Limitations and future research prospects are discussed in the next section.

As pointed out above, co-link count is a similarity measure which has been used in previous papers to evaluate competitive positions of companies within a single industry. Extending those earlier studies, the current study analyzed heterogeneous companies in terms of activity and, therefore, our first expectation is to find clusters of similar companies, that is, companies belonging to the same industry are expected to be grouped.

3.1. Dow Jones Industrial

Dow Jones Industrial is made of the top 30 companies in the USA. The composition at the time of the study and the industry affiliation of companies are in Appendix 1. Figure 1 is the MDS map for this stock. Only significant clusters are drawn to make the interpretation clearer.

Three major groups of companies are discernable:

- IT group includes the following companies: fixed line telecommunications (Verizon, AT&T), computer services and hardware (IBM, HP), semiconductors (Intel) and software (Microsoft). Positions of companies within this cluster are also meaningful. The three

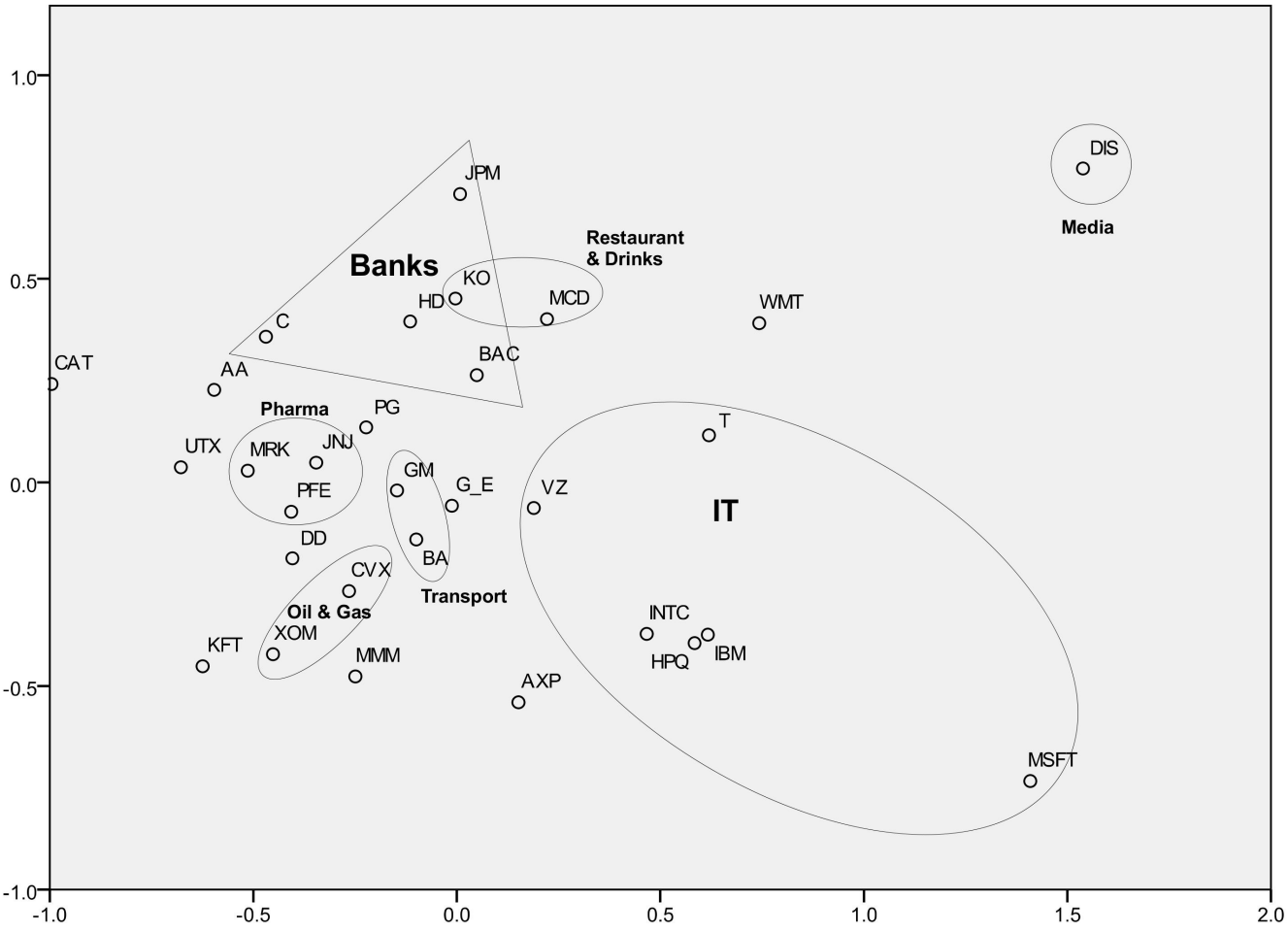
hardware companies are closer to each other and the same is true for the two telecommunications companies. Microsoft is the only software company here and it occupies a distinctive and domination market position, so it is distant from other companies.

- Financial group is made of banks and other financial services. There are three banks in the index Bank of America (BAC), Citigroup (C), JP Morgan (JMP) that are located within a triangle in the upper part of the map. Clustering is not perfect as two other companies appear in the same space. A fourth company to take into consideration is American Express that is placed down in the map closer to IT group. A common feature of financial companies in the study is that they seem to be located in a similar position as shown in Fig. 5 and 7 later.
- Other three clusters are well identified (pharmaceutical, oil & gas, transport). Scattered around these three clusters are other companies belonging to various other industries. A common feature of these companies and the companies in the three clusters is that they belong to traditional industries whose activities are no so focused on information.

The positions of two other companies are worth noting. As the only media conglomerate in Dow Jones, Disney does not compete directly with any other companies in the map. The nature of its activity, with most of its products and services likely to be digitalised, implies a strong presence on the Web which is also used as one of its main distribution channels. It received 6,070,000 external inlinks, second only to Microsoft (46,600,000). The fact that the company received millions of inlinks but had a very small percentage of co-links with other companies, together with the lack of competition with other companies in the index, explains Disney's outskirt position very well. Although there is no other media company in the map to form a cluster with Disney, it is circled in Fig. 1 for comparisons with media companies in other maps later. Like Disney, Wal-Mart is positioned as an outlier and attracted 1,210,000 external inlinks. It is worth to note how a retailer company as Wal-Mart received so many inlinks. This could be explain by the evolution in the business model that turn to be more focused on Internet selling and distribution of products. It is the only company other than the IT and Media ones that received more than 1,000,000 inlinks.

Regarding inlink counts, it is worth to note that companies on the right side of the map are the ones with more inlinks. All companies (except Verizon) received more than 1,000,000 inlinks. Companies with fewer inlinks (less than 150,000 inlinks) are located at the left side of the map. Companies in the middle of the map have mostly between 200,000 and 300,000 inlinks. It is perhaps not surprising that IT companies and consumer oriented companies (Disney and Wal-Mart) are more visible on the Web and therefore they attract more inlinks. A study of Chinese Web inlink patterns also found that Websites of consumer oriented companies received more inlinks (Vaughan, Tang & Du, 2009).

Figure 1. MDS map for Dow Jones Industrial



3.2. FTSE 100

FTSE 100 companies represent more than 80% of the market capitalization of the whole London Stock Exchange and constitutes the most widely used UK stock market index. Figure 2 shows the MDS map for the top 80 companies of the index in terms of inlinks. Although the high concentration of points in the center of the map makes interpretation unfeasible, this picture is especially useful as it shows clearly how some well defined groups are placed at the outskirts of the map. These clusters are IT, Media and Leisure companies. Because these IT companies are not as homogenous as the IT companies in the Dow Jones, they are not grouped into a single space; however, they are all located in a peripheral position of the map, as observed for Dow Jones. In this case, three out of the four most important media companies in the index are located in a common area. Only Pearson remains very close but out of this group; however its affiliation is much clearer in Fig. 3. Leisure group of companies is made of user oriented companies such as:

Carnival, Thomas Cook and Tui Travel (travel agents), and Intercontinental Hotels (hotels). Travel business is one of the activities that have implemented e-commerce intensively. Although not showed in the map, financial companies are relatively well clustered in an intermediate position in the right part of the cloud of points.

Figure 2. MDS map for FTSE 100 (Top 80 companies in terms of inlinks)

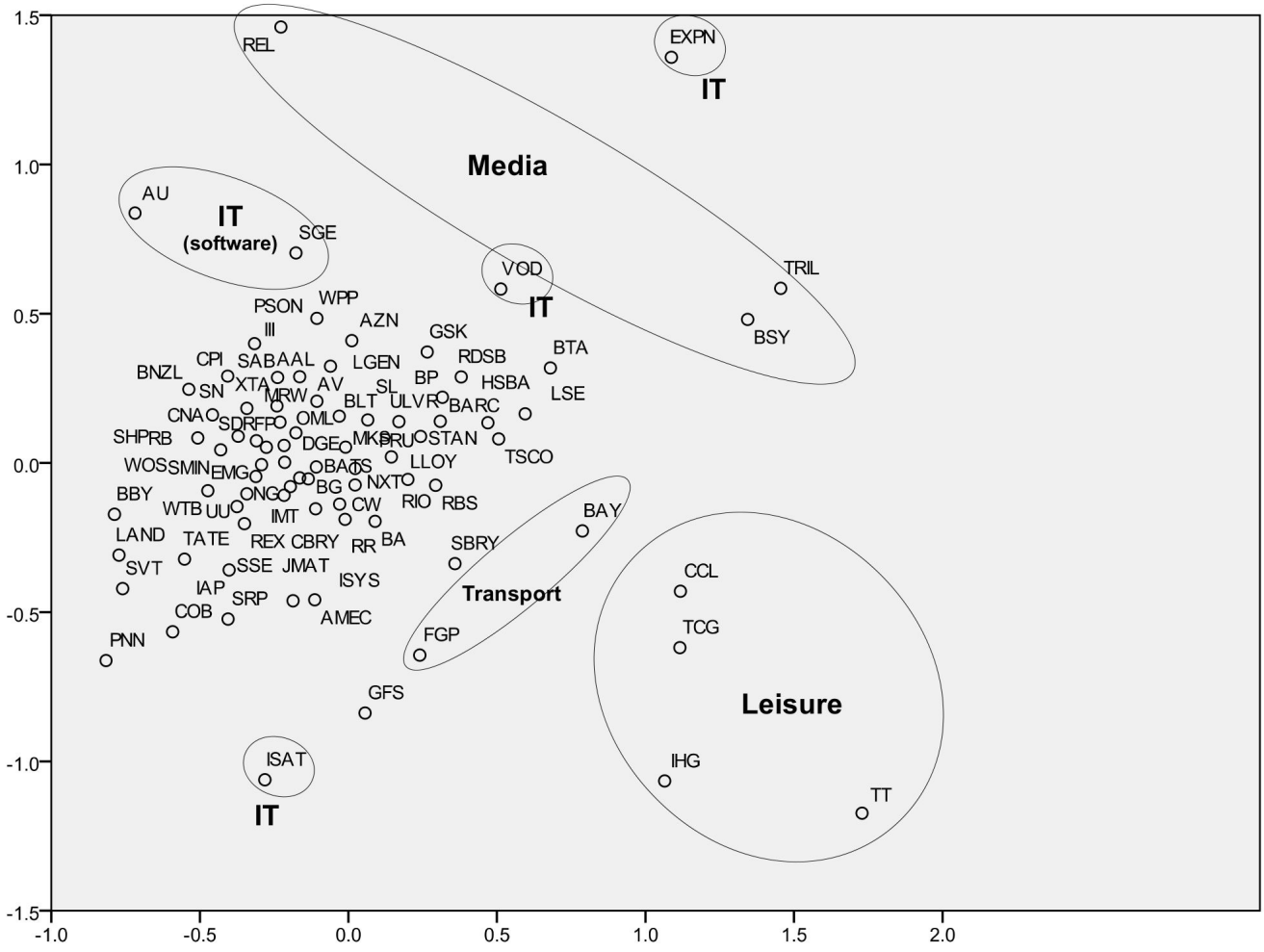
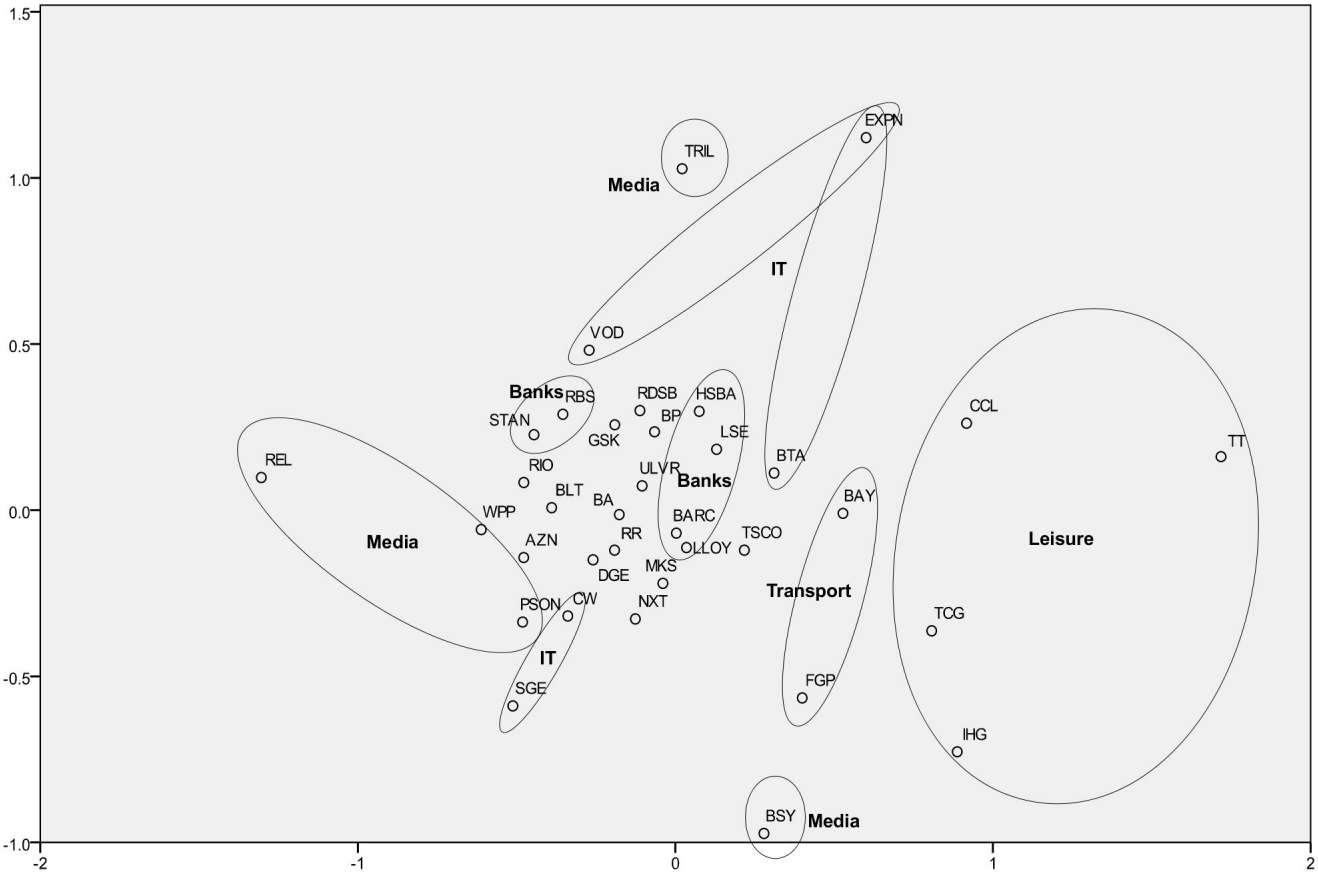


Figure 3 is the result for the data set of the top 35 FTSE companies in terms of inlink count. This selection probably overweighs some industries that by nature are likely to receive more inlinks (such as Media or IT). This approach provides an opportunity to test if more meaningful clusters will emerge from this microscopy approach. The resulting MDS map is a diverse representation of companies showing consistent clustering patterns that we have also found in other indexes. The results suggest that IT, Media and Leisure companies are placed at the outskirts of the map, with banks again placed adjacent to them. Clusters here are not clearly distinguished, probably due to

the heterogeneity of activities and the high number of companies included in the IT and Media groups. For instance, media companies include such a distinct businesses as TV, books, press or a marketing services company (not strictly a media but strongly connected).

Figure 3. MDS map for FTSE 100 (Top 35 companies in terms of inlinks)



3.3. Eurostoxx 50

Dow Jones Eurostoxx 50 is a stock index that intends to provide a blue-chip representation of Supersector leaders in the Eurozone. It is composed of 50 companies from seven countries: France (19), Germany (13), Netherlands (5), Italy (6), Spain (5), Finland (1) and Luxemburg (1).

This index is heterogeneous in terms of countries and industries. Therefore two different interpretations based on these criteria are developed. Firstly, Figure 4 shows the MDS map for Eurostoxx 50 according to the country location of the companies. Companies belonging to the same country appear to be together. This is especially clear for French companies. In the German space we can find a number of companies belonging to other countries (dashed line), but we will

see later that this could be explain in some cases by the strength of industrial pattern. Most Spanish companies are grouped very close. Dutch companies occupy a central position in the map between German and French companies. This position reflects the real geographical and economic situation in the Eurozone, where Dutch companies, due to the small size of its domestic market, tend to be more international. The only Luxembourgish company in this map, ArcelorMittal (ISPA), is embedded into the French space. This could be explained by the fact that Arcelor is the result of a merge in 2002 of three companies: Aceralia (Spain), Usinor (France) and Arbed (Luxembourg). This merger combined with the geographical proximity justifies its position.

Figure 4. MDS map for Eurostoxx 50, interpreted by countries.

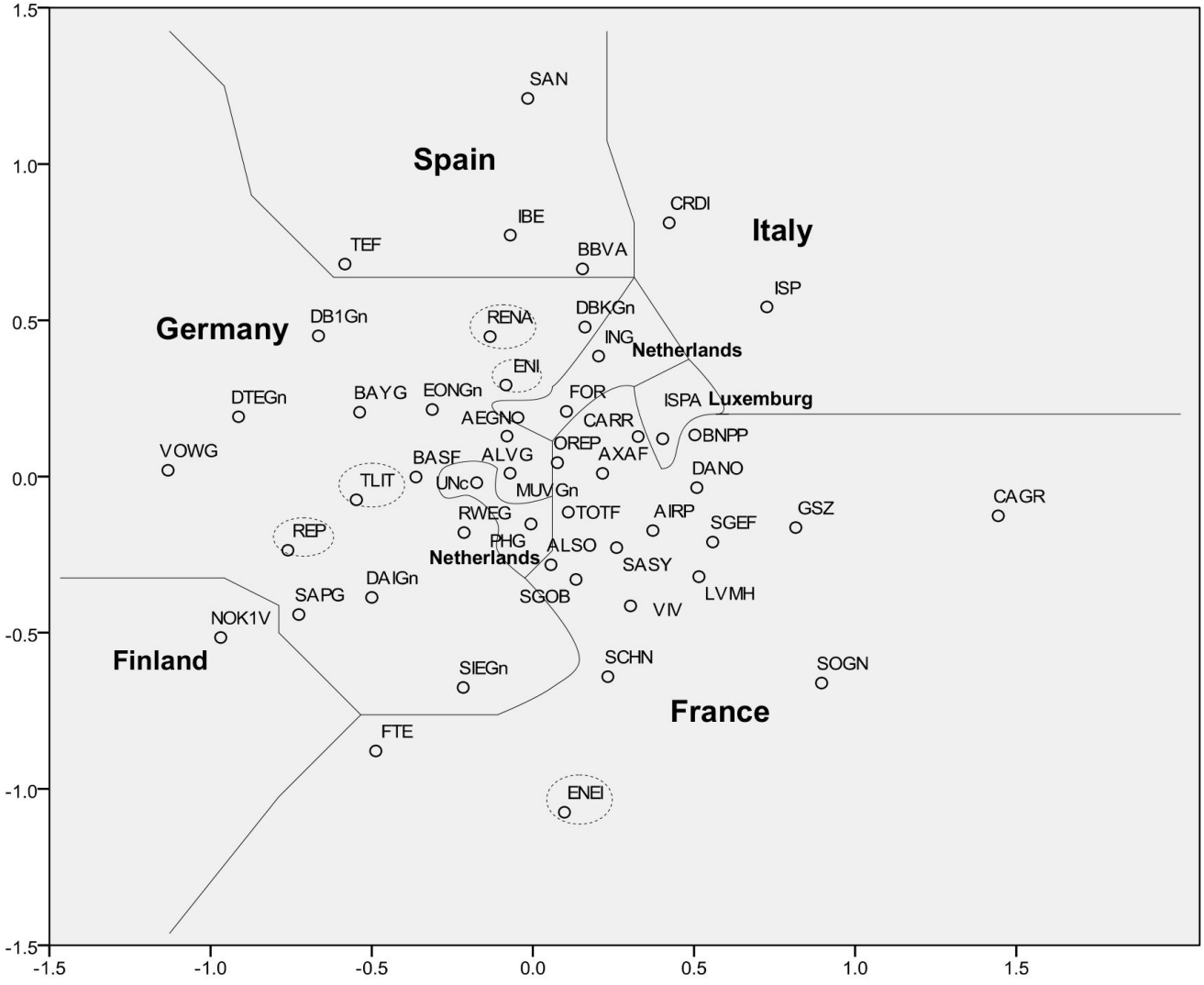
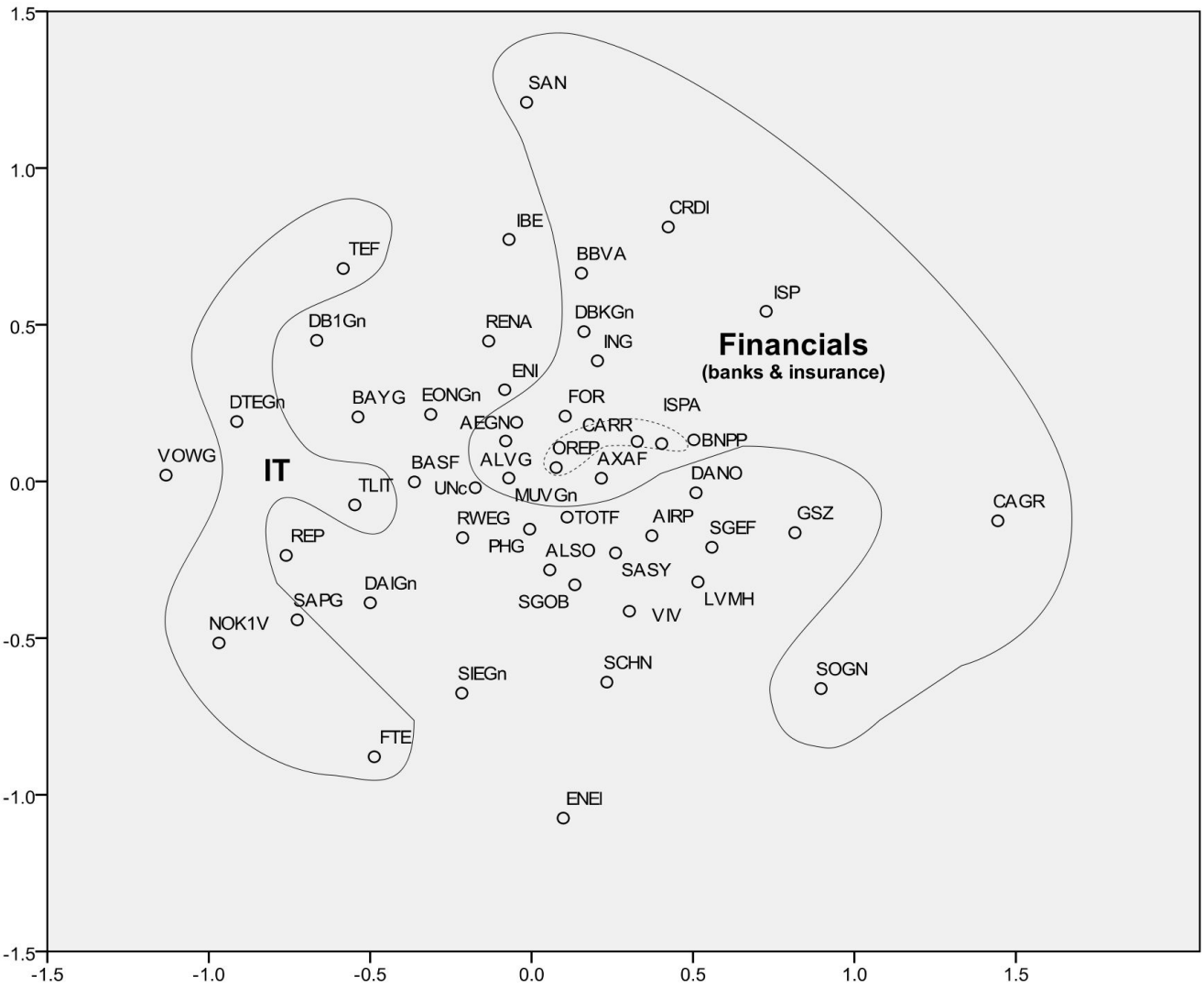


Figure 5 contains the same MDS map interpreted according to industries. Only two main groups are clearly identified: IT and Financial groups. No media or leisure companies are included in this

index. Financial group is mainly made of banks and insurance companies. All of them, with the exception of Deutsche Boerse (DB1Gn) that is close to IT companies, are together in a vast area regardless of country origin. There are also three companies (dashed line) from other industries that are inside of this cluster. IT companies are placed in the left part of the map including six companies from five countries (two of them are German). The existence of this cluster of IT companies explains why the German area of the map (Figure 4) presents some companies from other countries. IT has consistently immerged as a clearly defined cluster of companies in all stock exchanges in the study. This could be explained by the nature of its activity that is based intensively on information content and infrastructure. The financial companies in this map also form a clear cluster, probably due to its information based activity. The rest of the companies do not show a clear pattern in terms of industries, consistent with the results in other maps. Market integration in the Eurozone is likely to be deeper in these information sectors that are not restricted by geographical location for the provision of services and goods.

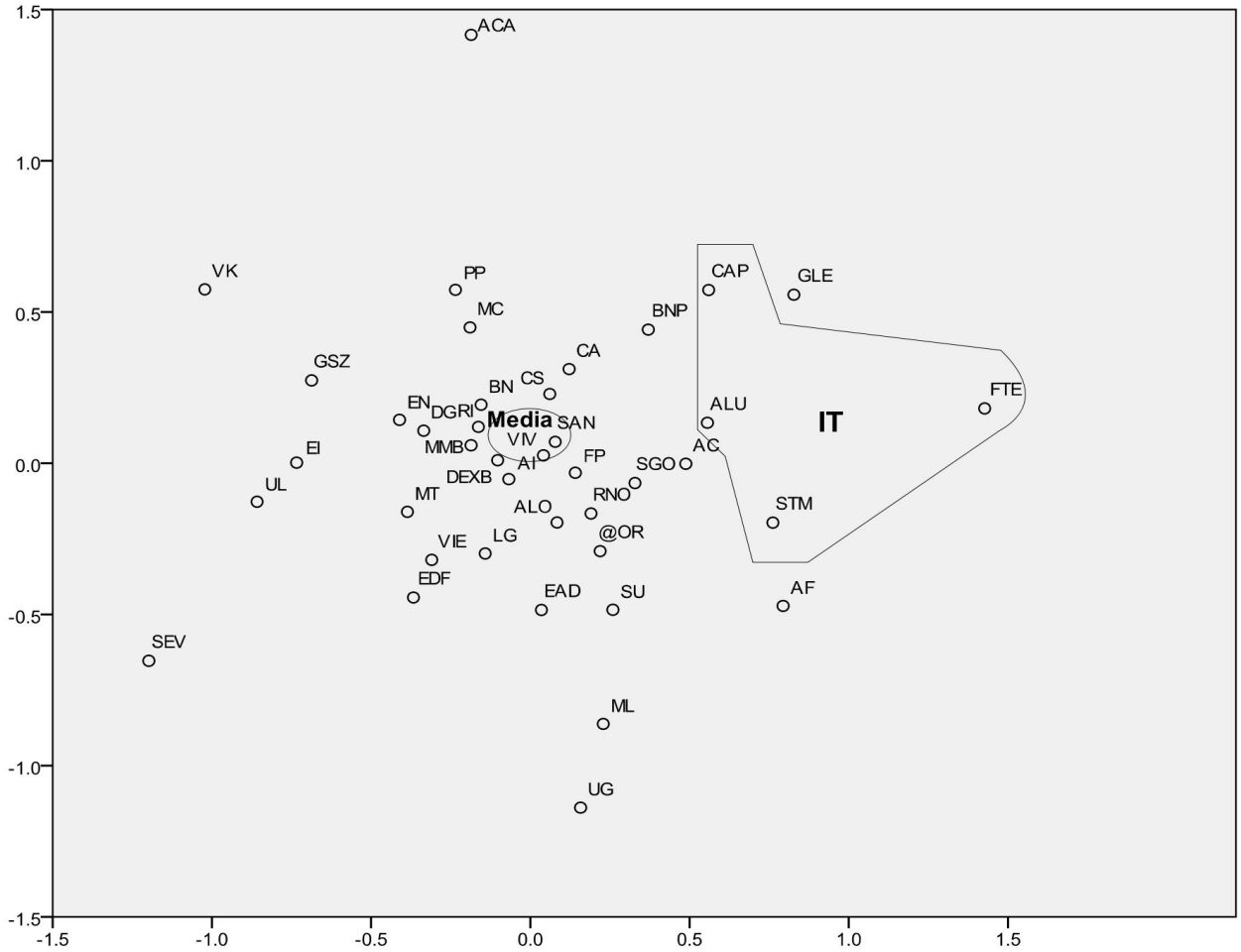
Figure 5. MDS map for Eurostoxx 50, interpreted by industries



3.4. CAC 40

The CAC 40 is a benchmark French stock market index that represents a capitalization-weighted measure of the 40 most significant values among the highest market caps on the Euronext Paris Bourse. Figure 6 is the MDS map for this data set. It shows that IT companies are in a cluster that is far removed from other companies. The position and the fact that these companies are top inlinked are coherent with findings in other indexes. Accor (AC), a hotel company (Leisure industry) receives a high number of inlinks and is placed very close to IT companies. Banks are more spread in the map but two of them are located near the IT companies: BNP Paribas (BNP) and Société Générale (GLE). Another bank, Crédit Agricole (ACA) is at the border of the map. However, a few companies in other industries present positions not expected at the outskirt of the map: Peugeot (UG, automobiles), Suez (SEV, waste & disposal services), Vallourec (VK, industrial machinery). However a result inconsistent with previous findings is that Vivendi (VIV), a media company, is placed in the middle of the map. Also Lagardere (MMB), a group with very diverse activities but with a significant publishing segment is also placed in the center very close to Vivendi. Moreover, inlink counts for these companies are significantly low in relation to other companies in different industries and especially with other media companies in other indexes. These unexpected results could partially be explained by the language problem. Further research on this particular market is necessary to get a better understanding of some particular companies.

Figure 6. MDS map for CAC 40

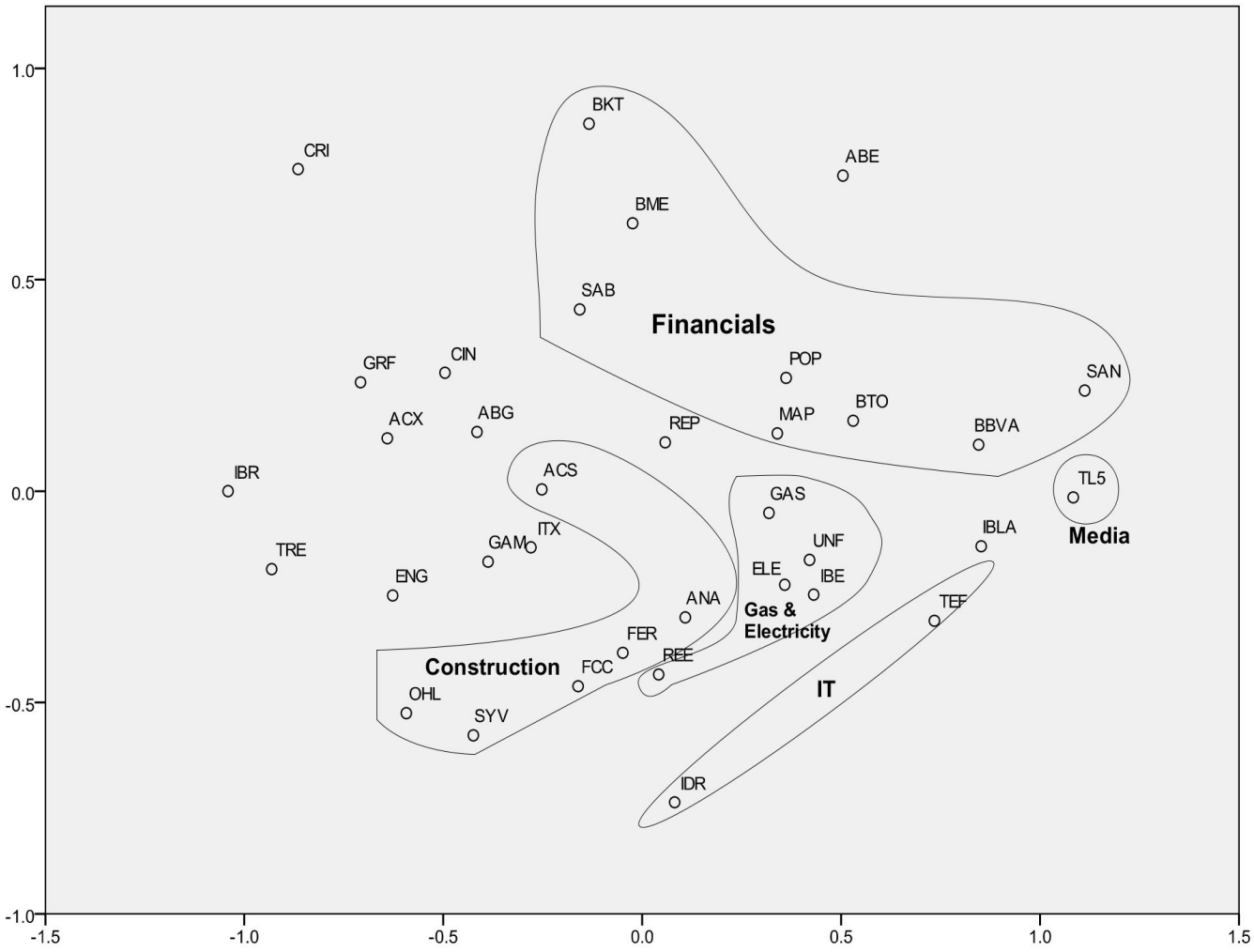


3.5. Ibex 35

The IBEX 35 is the benchmark stock market index of the Madrid Stock Exchange, comprising of the 35 most liquid Spanish stocks. Figure 7 confirms some of the patterns discovered in other stocks in this study.

The only media company TL5 is placed at the outskirt of the map. The cluster with the two IT companies is placed near it. Financial companies, including banks, insurance and traders, form a clear cluster. The rest of the companies in the center of the map belong to different industries with no clear pattern or not enough number of companies to form clusters. However, two of the key industries in Spain are clearly defined in the map: Construction and Gas & Electricity. Construction has been one of the leading industries in the country during the last decades and today it is the one that is suffering economic crisis more deeply.

Figure 7. MDS map for IBEX 35



4. Conclusions and future research

The originality of this study is based on the application of co-link analysis to heterogeneous companies listed in five major stock exchanges indexes. Previous co-link studies examined companies within a single industry with successful results in mapping competitive positions. However, due to the diverse industrial affiliation of the companies in this paper, the expectation is companies in the same industry to appear together in clearly identified clusters, according to the similarity principle used in previous research. Competitive analysis could also be developed but only within a previously identified cluster of homogeneous companies. However, this is not possible in this study due to the scarce number of companies belonging to the same industry. So far, studies based on heterogeneous types of organisations have only been developed in order to test the triple helix theory in the Web, but their perspective is not purely commercial or economic. Previous section examined the results individually, but there are some overall exploratory

conclusions that are promising in order to perform future research and to advance in the application of Webometric techniques to business studies.

Individual MDS maps show that the expectation of finding same industry companies grouped together is only partially fullfilled. Industries whose business model is more information centered form distinct clusters located at a position that is distant from other companies located in more central position. Information centered industries, according to our interpretation, comprise mainly four groups: companies based in the production of information contents (so called Media in the study), companies providing information infrastructures and services (IT), companies very intensive in applying e-commerce techniques (Leisure companies in the tourism industry) and Financial companies. Financial companies, including banks, insurance and other financial services, are information intensive in the sense that they do not usually provide physically tangible services or that they only operate in virtual venues (e.g. stock exchange companies, banks that are mainly present on the Internet, etc.).

In addition to their location in the MDS maps, media and IT companies receive more inlinks than any other companies that belong to more traditional industries. This particular inlink pattern reflects a business model that is highly exposed to Internet. This is true of financial companies to some extent. Generally speaking, they are second only to IT companies in the number of inlinks received. This explain why, in most countries, financial companies are located in a position between the center of the map, where more traditional companies are placed, and the IT and media companies. Therefore, generally speaking, companies that are located away from the centre of the map are the ones that depart from a traditional business model that is slower, by its own industrial nature or by a management decision, to move to an online context. This explanation is supported by the use of the Jaccard index in our data analysis. Because information centered companies receive many inlinks but relatively fewer of them are co-links with other companies, the Jaccard index scores tend to be lower. That is why they are positioned away from the center, at the outskirt of the map.

But why do different patterns exist based on information? The so called information centered companies have undergone a process of digitalization with the result of an increasing presence on the Web. The more dramatic changes are affecting the information content production industry (Media industry, e.g. press, music, cinema, and so on). These companies need to reinvent their business models in order to face the significant digitalization process of the last two decades. In most of the cases intellectual property issues are brought up. Companies that in the past seemed

to be stable and to enjoy dominant positions in the market have found themselves competing in a totally new and challenging economic and social environment.

Eurostoxx 50 is the only index in the study that is made of companies belonging to different countries. This fact allowed us to analyze economic integration in the Eurozone. Results show that interpretation by countries of Eurostoxx 50 map fits well the geographical dimension on the economies involved. Except for the information based industries (IT and Financial) no industrial clusters are clearly identified in the map. As a contrast, Dow Jones map is more clearly defined in terms of industries, underlining the deeper economic integration of the US market. This suggests that the Eurozone economy is not as tightly integrated yet. Perhaps future studies will reveal less country division as the trend of integration continues. An effective economic integration is likely to be easier in industries that are based on information technology. This integration is fostered by the rapid process of digitalization of EU countries. However, there could be some limitations in the Eurostoxx 50 co-link analysis due to language differences.

This paper reveals that Web co-link data constitute a readily available source of information to obtain new insights into the business environment. The study of heterogeneous companies belonging to different industries allows managers, financial analysts and other stakeholders to locate their company's activity in relation to others and to identify differing business modes and to follow their evolution over time. If a company is trying to increase their presence online, co-link analysis could be used by managers to monitor the progresses of their efforts by observing how the co-link MDS maps change. Therefore, co-link analysis could become a managerial tool useful for companies changing their IT strategies. Finally, this paper used online data to provide confirmatory evidence of the changes in business models of information centered industries due to the digitalization process of the last decades.

Conclusions are mainly exploratory but relevant enough to develop a whole set of future confirmatory studies and to open a new direction in the webometric research of commercial and economic issues. Additional evidence about the suggested explanation of different business models in terms of information is needed. Another important research direction to pursue is to monitor the changes of the MDS maps by collecting and analyzing data over time. Given the rapid development of the Internet in the last decade, it is likely that years ago IT and Media companies were not located in such an external position in the maps. Evolution in the coming years could provide confirmation about the transformation in the business models. Same longitudinal perspective could be applied to monitor changes in the Eurozone in order to find out whether country lines become less visible and industry clusters become clearer. Finally, a content analysis

study to examine the motivations of Webpage inlinking could be useful to determine if there are significant differences between information centered and traditional business model companies.

Although the paper explored the use of co-link analysis for heterogeneous groups of companies, providing new ways of studying commercial and economic issues, the study has some limitations, most of which are beyond our control. The interpretation of the results could differ due to factors such as different official industrial classifications, alternative criteria to include companies within a cluster, or specific knowledge of the researcher about a particular economic context. The industrial composition of the indexes differs due to the economic characteristics of each country. Also the number of companies in each index is not homogeneous. For example, FTSE 100 includes several leisure companies although Dow Jones Industrial does not. Sometimes only a single company is representative of an industry and no cluster can be identified. However the position of the particular company can be interpreted in comparison to similar companies in other indexes. Finally, language differences could affect the results due to a potential language bias of the search engine used for data collection, especially for Eurostoxx 50 that includes companies from different countries.

References

Björneborn, L. and Ingwersen, P. (2004), "Toward a basic framework for webometrics", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 14, pp. 1216–1227.

García-Santiago, L. and de Moya-Anegón, F. (2009), "Using co-outlinks to mine heterogeneous networks", *Scientometrics ,* Vol. 79 No. 3, pp. 681-702.

Google (2006), "Google SOAP Search API Reference", available at: http://www.google.com/apis/reference.html#2_2 (accessed 3 June 2009).

Google (2009), "Links to your site", available at: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=55281 (accessed 3 June 2009).

Heppes, D. and du Toit, A. (2009), "Level of maturity of the competitive intelligence function. Case study of a retail bank in South Africa", *Aslib Proceedings: New Information Perspectives*, Vol. 61 No. 1, pp. 48-66.

Kahaner, L. (1996), *Competitive Intelligence – How to Gather, Analyze, and Use Information to Move Your Business to the Top*, 7th ed., Touchstone, New York, NY.

Live Search (2007), "We are flattered, but...", available at:

http://www.bing.com/community/blogs/search/archive/2007/03/28/we-are-flattered-but.aspx (accessed 3 June 2009).

Porter, M. E. (1980), *Competitive Strategy: Techniques for Analyzing Industries and Competitors*, Free Press, New York, NY.

Porter, M. E. (1985), *The Competitive Advantage: Creating and Sustaining Superior Performance*, Free Press, New York, NY.

Reid, E. (2003), "Using Web link analysis to detect and analyze hidden Web communities", in Vriens, D. (Ed.), *Information and communications technology for competitive intelligence*, Ideal Group, Hilliard, OH, pp. 57-84.

Romero-Frías, E. and Vaughan, L. (2009), "Financial Distress of U.S. Banking Industry Viewed through Web Data", *12th International Conference on Scientometrics and Informetrics ISSI* 2009, Rio de Janeiro, Brazil, July 14-17, 2009.

Shaw, D. (2001), "Playing the links: interactivity and stickiness in .com and "not.com" websites", *First Monday*, Vol. 6 No, 3, available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/837/746 (accessed 10 May 2009).

Small, H. (1973), "Co-citation in the scientific literature: a new measure of the relationship between two documents", *Journal of the American Society for Information Science*, Vol. 24 No. 4, pp. 265-269.

Stuart, D. and Thelwall, M. (2006), "Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry", *Research Evaluation*, Vol. 15 No. 2, pp. 97–106.

Tan, B., Foo, S. and Hui, S. C. (2002), "Web information monitoring for competitive intelligence", *Cybernetics and Systems*, Vol. 33 No. 3, pp. 225-251.

Thelwall, M. and Wilkinson, D. (2004), "Finding similar academic Web sites with links, bibliometric couplings and colinks", *Information Processing and Management*, Vol. 40, pp. 94–101.

Vaughan, L. (2006), "Visualizing Linguistic and Cultural Differences Using Web Co-Link Data", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 9, pp. 1178-1193.

Vaughan, L. and Thelwall, M. (2004), "Search engine coverage bias: Evidence and possible causes", *Information Processing & Management*, Vol. 40 No. 4, pp. 693-707.

Vaughan, L., and You, J. (2006), "Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market", *Scientometrics*, Vol. 68 No. 3, pp. 611–628.

Vaughan, L., and You, J. (2008), “Content assisted web co-link analysis for competitive intelligence”, *Scientometrics*, Vol. 77 No. 3, pp. 433-444.

Vaughan, L., and You, J. (2009), “Keyword enhanced Web structure mining for business intelligence”, *Lecture Notes in Computer Science*, Vol. 4879, pp. 161–168.

Vaughan, L., and Wu, G. Z. (2004), “Links to commercial websites as a source of business information”, *Scientometrics*, Vol 60 No 3, pp. 487–496.

Vaughan, L., Gao, Y. and Kipp, M. (2006) “Why are hyperlinks to business Websites created? A content analysis”, *Scientometrics*, Vol. 67 No. 2, pp. 291–300.

Vaughan, L., Kipp, M. and Gao, Y. (2007), “Are co-linked business web sites really related? A link classification study”, *Online Information Review*, Vol. 31 No. 4, pp. 440–450.

Vaughan, L., Tang, J. and Du, J. (2009), “Examining the robustness of Web co-link analysis”, To appear in *Online Information Review*, 33, (5).

Zuccala, A. (2006), “Author Cocitation Analysis Is to Intellectual Structure A Web Colink Analysis is to...?”, *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 11, pp. 1487-1502.

Appendix 1. Dow Jones Industrial companies

Labels	Company	Main URL	Industry	Linkdomain count
MMM	3M Co.	http://www.3m.com	Diversified Industrials	251,000
AA	Alcoa Inc.	http://www.alcoa.com	Aluminum	42,900
AXP	American Express Co.	http://www.americanexpress.com	Consumer Finance	1,120,000
T	AT&T Inc.	http://www.att.com	Fixed Line Telecommunications	2,700,000
BAC	Bank of America Corp.	http://www.bankofamerica.com	Banks	364,000

BA	Boeing Co.	http://www.boeing.com	Aerospace	247,000
CAT	Caterpillar Inc.	http://www.cat.com	Commercial Vehicles & Trucks	227,000
CVX	Chevron Corp.	http://www.chevron.com	Integrated Oil & Gas	233,000
C	Citigroup Inc.	http://www.citigroup.com	Banks	128,000
KO	Coca-Cola Co.	http://www.coca-cola.com	Soft Drinks	214,000
DD	E.I. DuPont de Nemours & Co.	http://www.dupont.com	Commodity Chemicals	141,000
XOM	Exxon Mobil Corp.	http://www.exxonmobil.com	Integrated Oil & Gas	264,000
GE	General Electric Co.	http://www.ge.com	Diversified Industrials	262,000
GM	General Motors Corp.	http://www.gm.com	Automobiles	242,000
HPQ	Hewlett-Packard Co.	http://www.hp.com	Computer Hardware	2,900,000
HD	Home Depot Inc.	http://www.homedepot.com	Home Improvement Retailers	236,000
INTC	Intel Corp.	http://www.intel.com	Semiconductors	2,070,000
IBM	International Business Machines Corp.	http://www.ibm.com	Computer Services	3,830,000
JNJ	Johnson & Johnson	http://www.jnj.com	Pharmaceuticals	91,300
JPM	JPMorgan Chase & Co.	http://www.chase.com	Banks	232,000
KFT	Kraft Foods Inc. Cl A	http://www.kraft.com	Food Products	142,000
MCD	McDonald's Corp.	http://www.mcdonalds.com	Restaurants & Bars	636,000
MRK	Merck & Co. Inc.	http://www.merck.com	Pharmaceuticals	253,000
MSFT	Microsoft Corp.	http://www.microsoft.com	Software	45,500,000
PFE	Pfizer Inc.	http://www.pfizer.com	Pharmaceuticals	130,000
PG	Procter & Gamble Co.	http://www.pg.com	Nondurable Household Products	272,000
UTX	United Technologies Corp.	http://www.utc.com	Aerospace	42,800
VZ	Verizon	http://www22.verizon.com	Fixed Line	332,000

	Communications Inc.		Telecommunications	
WMT	Wal-Mart Stores Inc.	http://www.walmart.com	Broadline Retailers	1,310,000
DIS	Walt Disney Co.	http://www.disney.go.com	Broadcasting & Entertainment	5,500,000

Appendix 2. FTSE 100 (top 35 inlinked companies)

Labels	Company	Main URL	Industry	Linkdomain count
BSY	B Sky B Group	http://www.sky.com	Media; Television	2,040,000
BARC	Barclays	http://www.barclays.co.uk	Financials; Bank	62,500
BAY	British Airways	http://www.britishairways.com	Transports; Airlines	241,000
BT-A	BT Group	http://www.bt.com	Telecommunications	451,000
CW	Cable & Wireless	http://www.cw.com	Telecommunications	20,300
CCL	Carnival	http://www.carnival.com	Leisure; Travel Agents	69,800
EXPN	Experian	http://www.experian.com	Economic intelligence services	423,000
FGP	Firstgroup	http://www.firstgroup.com	Transports	30,600
HSBA	HSBC Hldg	http://www.hsbc.com	Financials; Bank	188,000
IHG	Intercont Hotels	http://www.intercontinental.com	Leisure; Accommodation	142,000
LLOY	Lloyds Banking Grp	http://www.lloydstsb.com	Financials; Bank	31,300
LSE	LSE Group	http://www.londonstockexchange.com	Financials; Stock Exchange	149,000
PSON	Pearson	http://www.pearson.com	Media; Books	22,100
REL	Reed Elsevier	http://www.reed-elsevier.com	Media; Press	828,000
RBS	Royal Bank Scotland Group	http://www.rbs.com	Financials; Bank	26,600
SGE	Sage Group	http://www.sage.com	IT; Software	73,000
STAN	Standard Chartered	http://www.standardchartered.	Financials; Bank	37,400

		com		
TCG	Thomas Cook Group	http://www.thomascook.com	Leisure; Travel Agents	52,000
TRIL	Thomson Reuters	http://www.reuters.com	Media; Specialist Retailing	10,400,000
TT	Tui Travel	http://www.tuitravelplc.com	Leisure; Travel Agents	92,600
VOD	Vodafone Group	http://www.vodafone.com	Telecommunications; Mobile	148,000
WPP	WPP	http://www.wpp.com	Services; Advertising	34,700

Appendix 3. Eurostoxx 50

Labels	Company	Main URL	Country	Industry	Linkdomain count
AEGN	Aegon	http://www.aegon.com	NL	Financials, Life Insurance	11,600
ALVG	Allianz	http://www.allianz.com	DE	Financials, Insurance	38,600
AXAF	Axa	http://www.axa.com	FR	Financials, Insurance	32,300
BBVA	Banco Bilbao Vizcaya Argentaria	http://www.bbva.com	ES	Financials, Bank	33,800
SAN	Banco Santander	http://www.santander.com	ES	Financials, Bank	366,000
BNPP	BNP Paribas	http://www.bnpparibas.com	FR	Financials, Bank	112,000
CAGR	Crédit Agricole	http://www.credit-agricole.com	FR	Financials, Bank	26,200
DBKGn	Deutsche Bank	http://www.db.com	DE	Financials, Bank	76,900
DB1Gn	Deutsche Boerse	http://www.deutsche-boerse.com	DE	Financials, Investment Services	88,800
DTEGn	Deutsche	http://www.telekom.de	DE	Telecommunications,	222,000

	Telekom			Mobile	
FOR	Fortis	http://www.fortis.com	NL	Financials, Bank	21,800
FTE	France Telecom	http://www.francetelecom.com	FR	Telecommunications, Fixed Line	885,000
SOGN	Société Générale	http://www.societegenerale.fr	FR	Financials, Bank	52,800
ING	ING grp	http://www.ing.com	NL	Financials, Insurance	57,400
ISP	Intesa Sanpaolo	http://www.intesasanpaolo.com	IT	Financials, Bank	40,900
MUVGn	Muenchener Rueck	http://www.munichre.com	DE	Financials, Insurance	8,870
NOK1V	Nokia	http://www.nokia.com	FI	Technology, Technology Hardware & Equipment	1,340,000
TLIT	Telecom Italia	http://www.telecomitalia.com	IT	Telecommunications, Fixed Line	5,140
TEF	Telefonica	http://www.telefonica.es	ES	Telecommunications, Fixed Line	279,000
CRDI	Unicredit	http://www.unicreditgroup.eu	IT	Financials, Bank	58,000

Appendix 4. CAC 40

Labels	Company	Main URL	Industry	Linkdomain count
AC	Accor	http://www.accor.com	Leisure; Travel; Hotels	75,900
ALU	Alcatel-Lucent	http://www.alcatel-lucent.com	Technology Hardware & Equipment	69,100
BNP	BNP Paribas	http://www.bnpparibas.com	Financials, Bank	111,000
CAP	Cap Gemini	http://www.capgemini.com	Software & Computer Services	77,900
ACA	Crédit Agricole	http://www.credit-	Financials, Bank	26,300

		agricole.com		
FTE	France Teleco	http://www.francetelecom.com	Fixed Line Telecommunications	881,000
MMB	Lagardere	http://www.lagardere.com	Media; others	5,890
UG	Peugeot	http://www.peugeot.com	Automobiles	47,200
GLE	Société Générale	http://www.societegenerale.fr	Financials, Bank	51,500
STM	STMicroelectronics	http://www.st.com	Technology Hardware & Equipment	75,000
SEV	Suez Environnement	http://www.suez-environnement.com	Industrial Goods & Services	21,800
VK	Vallourec	http://www.vallourec.com	Industrial Goods & Services	2,600
VIV	Vivendi	http://www.vivendi.com	Media	23,500

Appendix 5. IBEX 35

Labels	Company	Main URL	Industry	Linkdomain count
ANA	Acciona	http://www.acciona.es	Construction	23,700
ACS	Actividades Construcciones y Servicios	http://www.grupoacs.com	Construction	3,710
BBVA	Banco Bilbao Vizcaya Argentaria	http://www.bbva.es	Bank	214,000
SAB	Banco de Sabadell	https://www.bancsabadell.com	Bank	5,860
BTO	Banco Español de Credito	http://www.banesto.es	Bank	33,200
POP	Banco Popular Español	http://www.bancopopular.es	Bank	18,200
SAN	Banco Santander Central Hispano	http://www.santander.com	Bank	367,000
BKT	Bankinter	https://www.bankinter.com	Bank	8,490
BME	Bolsas y Mercados Españoles	http://www.bolsasymercados.es	Investment services	15,800
MAP	Corporacion Mapfre	http://www.mapfre.com	Insurance	24,400
ELE	Endesa	http://www.endesa.es	Utilities	61,500
FER	Ferrovial	http://www.ferrovial.es	Construction	13,100
FCC	Fomento de Con. y Contratas	http://www.fcc.es	Construction	20,600
GAS	Gas Natural sdg	http://www.gasnatural.com	Utilities	28,800
TL5	Gestevision Telecinco	http://www.telecinco.es	Media & Advertisement	378,000
IBE	Iberdrola	http://www.iberdrola.es	Utilities	50,600
IDR	Indra, serie A	http://www.indra.es	Electronics &	37,700

			Software	
OHL	Obrascon Huarte Lain	http://www.ohl.es	Construction	1,520
REE	Red Electrica de España	http://www.ree.es	Utilities	19,900
SYV	Sacyr Vallehermoso	http://www.gruposyv.com	Construction	6,330
TEF	Telefonica	http://www.telefonica.es	Telecommunications	279,000
UNF	Union Fenosa	http://www.unionfenosa.es	Utilities	18,400

3.2.2 Financial Distress of U.S. Banking Industry Viewed through Web Data

The following study has been presented as a conference paper. This is the reference:

> ROMERO-FRÍAS, E. & VAUGHAN, L. (2009) "Financial distress of U.S. banking industry viewed through Web data". In *Proceedings of ISSI 2009 – the 12th International Conference of the International Society for Scientometrics and Informetrics* (pp. 206-210), Rio de Janeiro, Brazil, July 14-17, 2009.

Financial Distress of U.S. Banking Industry Viewed through Web Data

Abstract

Building on previous studies on inlink and co-link research on commercial websites, the current study attempted to apply and combine both methods to investigate the recent financial crisis in the U.S. banking industry. The methods combine Web content mining and Web structure mining. Two sets of inlink and co-link data were collected, one with the keywords “crisis”, “bailout” and “subprime” and one without any keyword. The one with the keywords was meant to include only inlink webpages that are likely to be about the financial crisis. The number of inlink pages that contained the keywords correlates significantly with the degree of a bank’s crisis measured by the amount of government bailout money. Both sets of co-link data were analysed using multidimensional scaling to generate maps of business competition. The comparison between the maps with and without keywords shows that the map generated from the co-link data with keywords depicts a more accurate image of the industry by clustering banks with more financial problems together.

1. Introduction

This paper reports an ongoing study that explores the possibility of using Web data to discover timely business information. Many economic and financial data such as revenue and profit are not timely as they are calculated quarterly or annually. In contrast, the Web is constantly changing. The Web data thus have the potential to provide more timely information. Unlike the established economic and financial variables, however, the meanings of usefulness of Web data are not very clear. Many studies have examined various types of Web data in terms of their relationship with economic and financial data. For example, Tumarkin & Whitelaw (2001) studied the relationship between Internet message board activity and abnormal stock returns and trading volume. Using social network analysis, Das and Sisk (2005) analyzed messages posted to stock boards. Jin, Matsuo & Ishizuka (2009) also used social network analysis but collected data from the general Web to rank companies. Findings from these studies contributed to our understanding on how to use Web data to gain business information but many more studies are needed for a clearer and firmer understanding. Toward this end, we are studying how Web data can be used to analyze the recent crisis of the U.S. banking industry. Specifically, we try to find out if the inlink and keyword data can be a measure of the degree of crisis, if the data can be used to visualize the industry in that banks with more distress are grouped together. We also collected various financial and economic data to triangulate the findings from the Web data. The research is currently in progress. This paper reports preliminary findings.

Previous research in Webometrics has analyzed commercial websites in order to explore new sources of business information that could be useful for data mining and business intelligence purposes (Thelwall, Vaughan & Björneborn, 2005). Link analysis has been used specifically for competitive intelligence (Chau, Shiu, Chan & Chen, 2007; Reid, 2003).

A recent study (Vaughan & You, 2008) proposed a method that combines page content with co-link data to achieve a more detailed picture of the competitive landscape of a sector within an industry. We apply this content assisted method to the banking industry in order to study the impact of the current financial and economic crisis. The keywords selected to refine the search are “crisis”, “bailout” and “subprime”, which are used frequently to describe the current worldwide financial problems. These keywords are intended to filter out pages that link the banks, whether they are inlinks or co-inlinks, for reasons other than the financial crisis. From a methodological point of view, this method provides new possibilities to use Webometric techniques in the field of commercial websites to research specific issues. We supported our co-link analysis with correlation test between inlinks and data on the degree of financial crisis. Definitions of inlink and co-link follows that by Björneborn & Ingwersen (2004).

2. Methodology

The bank sector was selected because it is at the origin of the financial crisis worldwide, specially in the United States. The crisis in this industry has shown particular characteristics, coming from the subprime mortgage crisis to the government rescue plans to save the industry. Moreover, statistics about the use of the Web by companies usually rank financial industry at the top, only below the information technology industry, which has been already studied. These features make the banking industry an appropriate and relevant subject to combine the aforementioned Webometric techniques.

Our first attempt to study the crisis focused on a set of international top banks. However, this approach faced some limitations, such as the different home languages of the companies that could bias the collection of data using specific keywords. The varied performance of national economies was also a significant deterrent because the heterogeneous conditions made it more complicated to analyse the impact of financial crisis in the companies. Many countries have also established rescue plans to support banks, but the conditions applied for the allocation of funds and the amounts were not comparable and these data were not publicly available for all countries.

Therefore, we decided to focus in the U.S. banking sector and more specifically in the banks listed in the New York Stock Exchange (NYSE). In the U.S., the Federal Government designed an unprecedented and expensive rescue plan to buy financial assets of the banks in order to protect a general bankruptcy of the financial system. In order to have more homogeneous data and to select a collection of major U.S. banks, we took the 46 companies included in the NYSE (www.nyse.com). These companies are listed on NYSE webpage under the following labels: Industry-Financials, Supersector-Banks, Sector-Banks, and Subsector-Banks. Only U.S. national companies with common stock traded were taken into account (the list of companies was taken on 16th January 2009, see Appendix 1). Additionally, Wachovia Corporation, that was not in the list due to the recent merge with Wells Fargo, was included.

Different measures of the impact of the financial crisis in the banks were considered, e.g. the change in the banks' stock prices over a one-year period. After exploring various measures with publically available data, we decided that the most reliable measure of a bank's of financial distress is the amounts of money it received under the Troubled Asset Relief Program (TARP) established by the U.S. Federal Government. Data were collected from a Special Report by CNN Money

(2009). Only 22 out of the 47 banks in the study had received money under the TARP.

Yahoo! is used for data collection as the other two major search engines in the market, Google and Live Search (MSN) do not allow to perform the type of queries shown in Table 1.

Table 1. Yahoo! Query Syntax for inlink and co-link data.

Type of data collected	Yahoo! Query
Inlinks without keywords	linkdomain:boh.com –site:boh.com
Inlinks with keywords	linkdomain:boh.com –site:boh.com (crisis OR bailout OR subprime)
Co-links without keywords	(linkdomain:boh.com -site:boh.com) AND (linkdomain:cnb.com -site:cnb.com)
Co-links with keywords	(linkdomain:boh.com -site:boh.com) AND (linkdomain:cnb.com -site:cnb.com) (crisis OR bailout OR subprime)

“Linkdomain” command searches for webpages that link to all pages of a site, while “link” command searches for pages linking only to a particular URL. Both commands were tested in the study. Data collected with the linkdomain command generated MDS maps that clustered troubled banks closer. Four banks had to be excluded from the co-link analysis because they did not have any co-link with other banks in the study. If we used the “link” command instead of the “linkdomain” command, more banks may have to be excluded because of the lack of co-links with other banks. Appendix 1 shows a sample of banks included in the study. The complete list of banks is omitted due to the space limitation.

3. Preliminary Results / Findings

There are significant ($p<0.01$) correlations between the number of inlinks to company websites and the amount of bailout money received. Correlation coefficients are 0.72 and 0.78 for the queries without and with keywords respectively. Both correlation coefficients are high but the one with the keywords is higher, suggesting that using the keywords in the search did add some information about the crisis. However, the difference between the two is not very large. Further research is needed to explore how the search can be revised to hopefully retrieve data with more crisis information.

MDS maps obtained from the two sets of co-link data are shown in Figure 1 and Figure 2. Figure 1 was generated from the data without the keywords while Figure 2 was from the data with the keywords. Small circles in black colour are used to mark the banks that received money. The strength of the black colour represents the amounts received.

Figure 1. MDS map without keywords (January 2009)

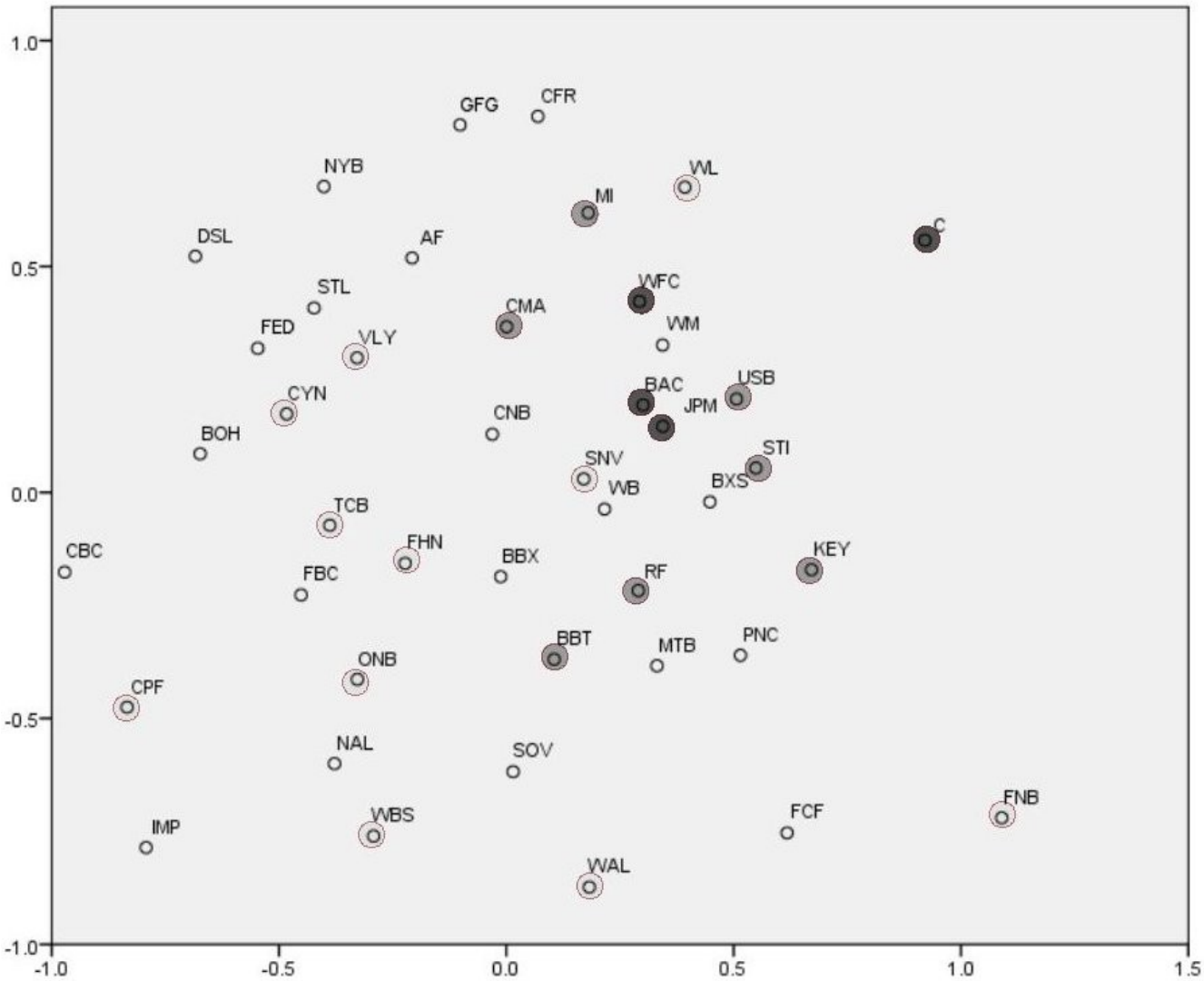
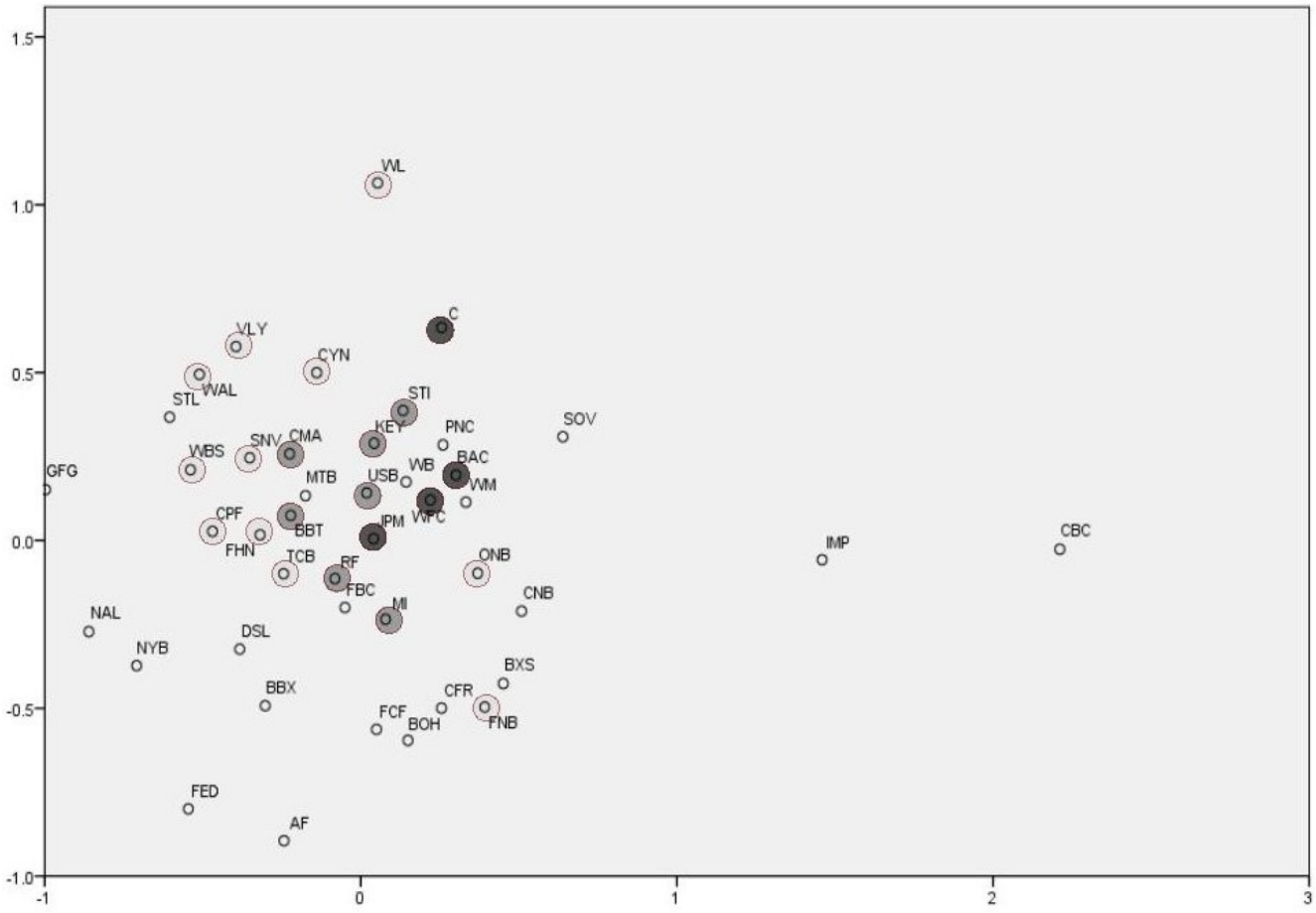


Figure 2. MDS map with keywords (January 2009)



Four banks in dark black colour received more than ten billion dollars, the seven median coloured banks received between one and ten billion dollars and, finally, banks in light black colour received less than one billion dollars. Comparing the two MDS maps, we can see that map with keywords (Figure 2) clusters the banks in approximately three layers with darker coloured ones in the centre, followed by median coloured in the outer layer and then the light coloured ones near the edge. Assuming that the government bailout money is approximately proportional to the degree of crisis of the affected banks, the map positioned banks according to their crisis level with the more affected ones in the centre and less affected ones on the outside. In contrast, MDS map without keywords (Figure 1) does not present a clear pattern that reflects the degree of crisis (banks more affected by the crisis are distributed in different parts of the map). This suggests that incorporating keywords in the inlink search helps to obtain specific information related to the keyword, in this case, the crisis. It makes sense that more co-links with the keywords would appear among banks that are in a similar degree of crisis because the more a bank is in crisis, the more pages discussing the crisis would link to it, i.e. the bank would receive more inlink pages that contained the crisis related keywords.

4. Extension [not in original conference paper]

In August 2009, a new round of data was collected to check the evolution in the maps. The analysis followed the same criteria as in figures 1 and 2. The results did not show any clear pattern of changes.

Figure 3. MDS map without keywords (August 2009)

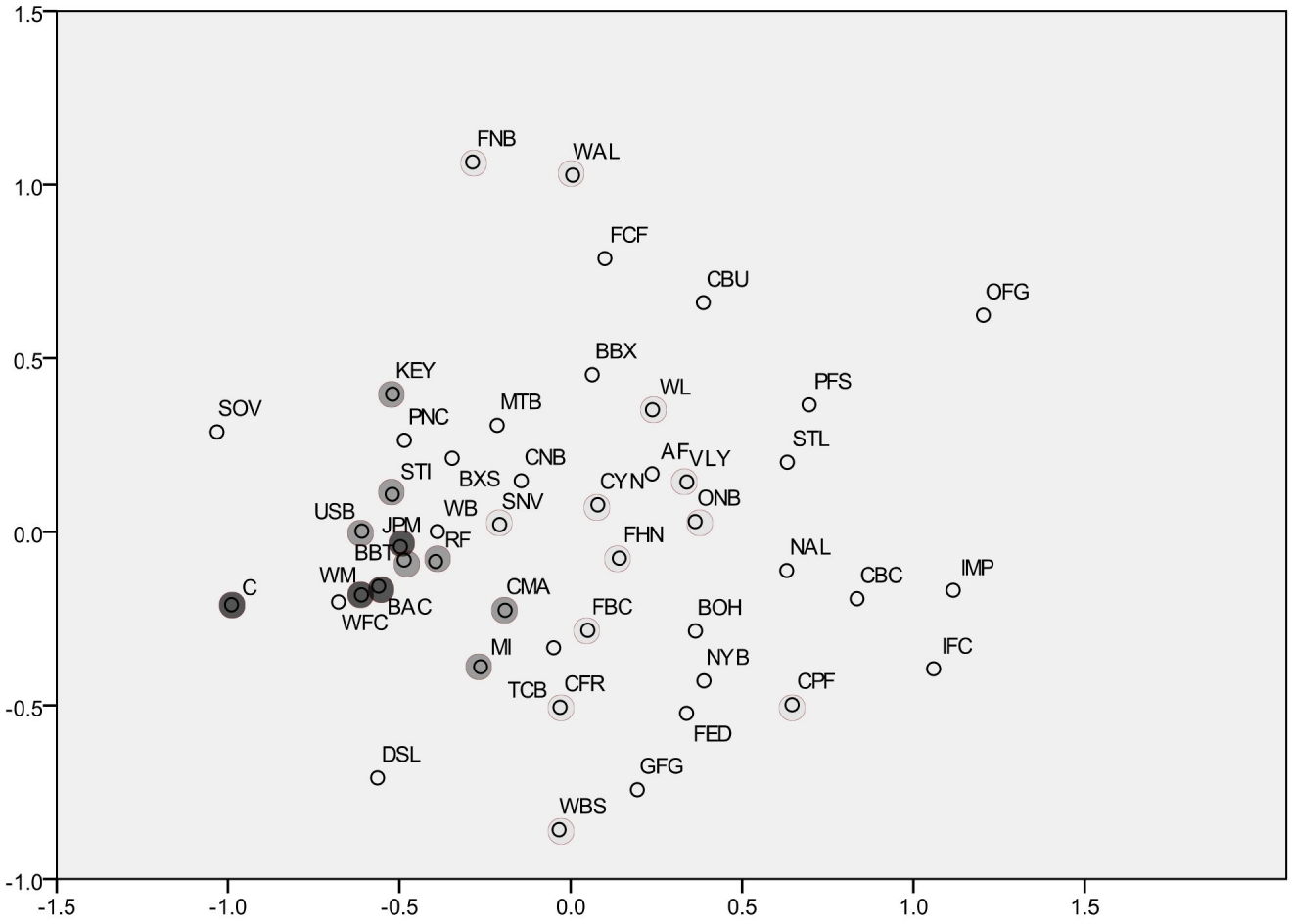
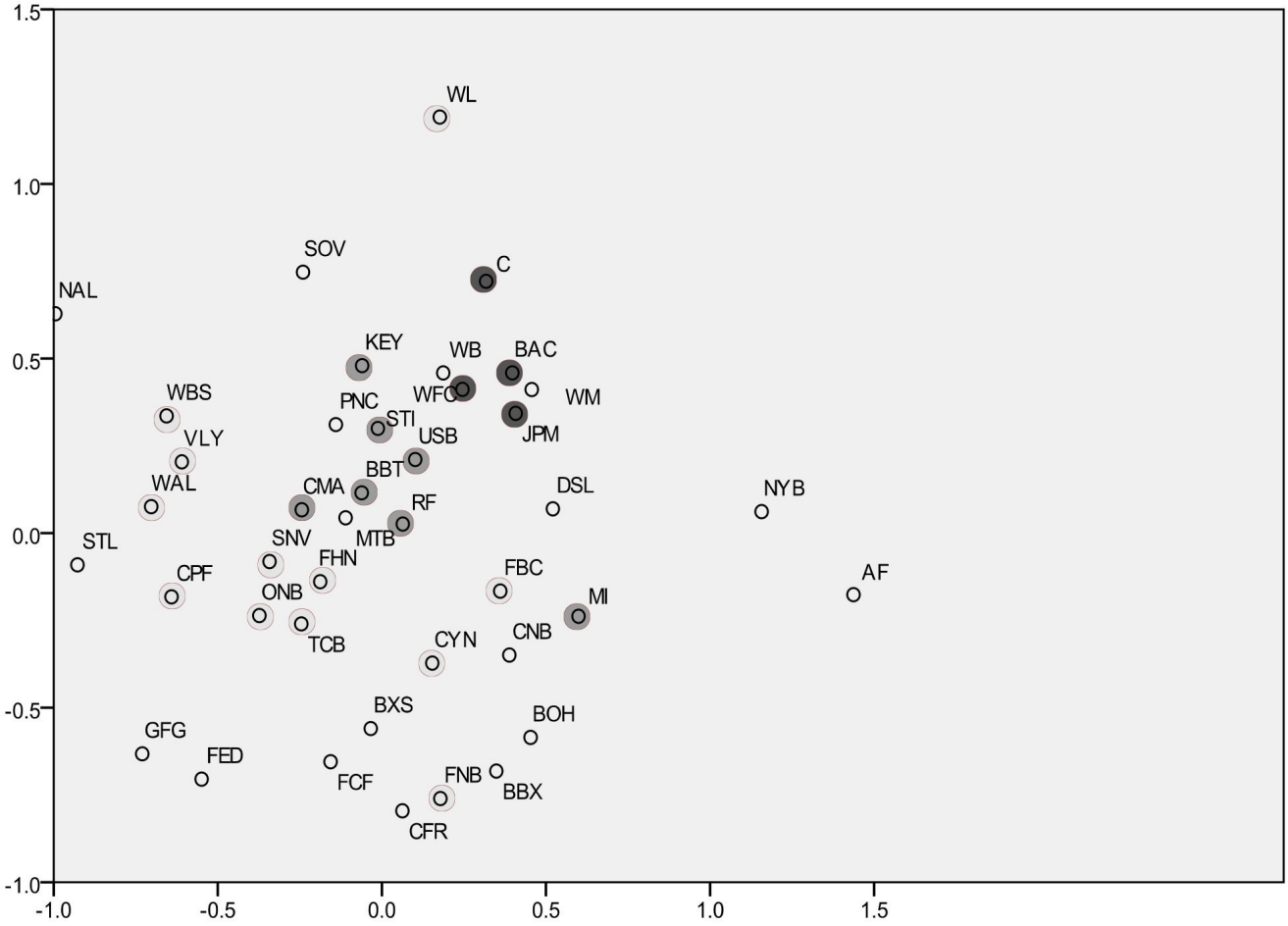


Figure 4. MDS map with keywords (August 2009)



5. Limitations [not in original conference paper]

The results obtained in August 2009 does not reflect market condition change. Some of the limitations of this approach that could explain this situation could be the following:

- Co-link analysis could not be an appropriate method to achieve the goals of the study. Not all the Web pages linking to a bank with financial distress link simultaneously to another bank in trouble. Therefore a direct link analysis could be a better approach.
- Also the variable used to measure the financial crisis is affected by a size effect because largest banks are the banks that will receive more money in absolute terms and therefore it is possible that we are only measuring the size of the entity instead of the intended variable.

6. References

Björneborn, L. & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14), 1216-1227.

Chau, M., Shiu, B., Chan, I. & Chen, H. (2007). Redips: Backlink search and analysis on the Web for business intelligence analysis, *Journal of the American Society for Information Science and Technology*, 58(3), 351-365.

CNN Money (2009), Economy rescue: Adding up the dollars. Retrieved Jan. 21, 2009 from http://money.cnn.com/news/specials/storysupplement/bailout_scorecard/index.html.

Das, S. R. & Sisk, J. (2005). Financial communities. *Journal of Portfolio Management*, 31(4), 112-123.

Jin, Y., Matsuo, Y. & Ishizuka, M. (2009). Ranking companies on the Web using Social Network Mining. In Ting, I. H. &Wu, H. J. (Eds). Web Mining Application in E-commerce & E-services. Berlink: Springer-Verlag, pp. 137-151.

Reid, E. (2003) 'Using Web link analysis to detect and analyze hidden Web communities', in: Vriens, D. (Ed.). *Information and Communications Technology for Competitive Intelligence*, pp 57-84, Hilliard, Ohio: Ideal Group Inc.

Thelwall, M., Vaughan, L. and Björneborn, L. (2005) 'Webometrics', in Cronin B. (ed.), *Annual review of information science and technology*, Vol. 39, pp 81-135, Medford, NJ: Information today.

Tumarkin, R. & Whitelaw, R. F. (2001). News or noise? Internet postings and stock prices. *Financial Analysts Journal*, 57(3), 41-51.

Vaughan, L. and You, J. (2008) 'Content assisted web co-link analysis for competitive intelligence', *Scientometrics*, 77 (3), 433-444.

Appendix 1. Banks included in the study

Company	Labels	URL	Money received under TARP 20/8/2009
Bank of America Corporation	BAC	http://www.bankofamerica.com	25.000.000.000
Citigroup Inc.	C	http://www.citigroup.com	25.000.000.000
JPMorgan Chase & Co.	JPM	http://www.chase.com	25.000.000.000
Wells Fargo & Co.	WFC	http://www.wellsfargo.com	25.000.000.000
US Bancorp	USB	http://www.usbank.com	6.599.000.000

Suntrust Banks Inc.	STI	http://www.suntrust.com	4.850.000.000
Regions Financial Corporation	RF	http://www.regions.com	3.500.000.000
BB&T Corporation	BBT	http://www.bbt.com	3.133.640.000
KeyCorp	KEY	http://www.key.com	2.500.000.000
Comerica Inc.	CMA	http://www.comerica.com	2.250.000.000
Marshall & Ilsley Corporation	MI	http://www.mibank.com	1.715.000.000
Synovus Financial Corp.	SNV	http://www.synovus.com	967.870.000
First Horizon National Corporation	FHN	http://www.firsthorizon.com	866.540.000
City National Corporation	CYN	http://www.cnb.com	400.000.000
Webster Financial Corporation	WBS	http://www.websteronline.com	400.000.000
TCF Financial Corporation	TCB	http://www.tcfbank.com	361.172.000
Wilmington Trust Corporation	WL	http://www.wilmingtontrust.com	330.000.000
Valley National Bancorp	VLY	http://www.valleynationalbank.com	300.000.000
* Flagstar Bancorp, Inc.	FBC	http://www.flagstar.com	266.657.000
Western Alliance Bancorporation	WAL	http://www.westernalliancebancorp.com	140.000.000
Central Pacific Financial Corp.	CPF	http://www.centralpacificbank.com	135.000.000
F.N.B. Corporation	FNB	http://www.fnb-online.com	100.000.000
Old National Bancorp	ONB	http://www.oldnational.com	100.000.000
Astoria Financial Corporation	AF	http://www.astoriafederal.com	0
BancorpSouth Inc	BXS	http://www.bancorpsouthonline.com	0
Bank of Hawaii Corporation	BOH	http://www.boh.com	0
BankAtlantic Bancorp, Inc.	BBX	http://www.bankatlantic.com	0
Capitol Bancorp Limited	CBC	http://www.capitolbancorp.com	0
Colonial Bancgroup Inc.	CNB	http://www.colonialbank.com	0
Community Bank System, Inc.	CBU	http://www.communitybankna.com	0
Cullen/Frost Bankers, Inc.	CFR	http://www.frostbank.com	0
Downey Financial Corporation	DSL	http://www.downeysavings.com	0
First Commonwealth Financial Corporation	FCF	http://www.fcbanking.com	0
Firstfed Financial Corporation	FED	http://www.firstfedca.com	0

Guaranty Financial Group Inc.	GFG	http://www.guarantygroup.com	0
Imperial Capital Bancorp, Inc.	IMP	http://www.itlacapital.com	0
Irwin Financial Corporation	IFC	http://www.irwinfinancial.com	0
M&T Bank Corporation	MTB	http://www.mandtbank.com	0
New York Community Bancorp, Inc.	NYB	http://www.mynycb.com	0
NewAlliance Bancshares, Inc.	NAL	http://www.newalliancebank.com	0
Oriental Financial Group, Inc.	OFG	http://www.orientalonline.com	0
PNC Financial Services Group,The	PNC	http://www.pnc.com	0
Provident Financial Services, Inc.	PFS	http://www.providentnj.com	0
Sovereign Bancorp, Inc.	SOV	http://www.sovereignbank.com	0
Sterling Bancorp	STL	http://www.sterlingbancorp.com	0
Wachovia	WB	http://www.wachovia.com	0
Washington Mutual Inc.	WM	http://www.wamu.com	0
* In the period from February to August only Flagstar Bancorp, Inc. received new funds. Also, some banks gave funds back.			

3.3 Combined analysis

3.3.1 A Webometric analysis of the international banking industry

A previous version of this study was presented as a poster in a conference. This is the reference:

> ROMERO FRÍAS, E. & VAUGHAN, L. (2009) “A Webometric Analysis of the Global Banking Industry ”. In *Proceedings of the 35th EIBA Annual Conference*, Valencia, Spain, December 13-15, 2009.

Web hyperlink patterns and the financial variables of the global banking industry

Abstract

This study combines both the inlink and co-link analysis techniques to make a complete examination of the international banking industry. Results from the analysis were compared with real financial situation of these banks to determine the validity and reliability of the link analysis methods. Top 50 international banks from 15 different countries were candidates for the study. Web hyperlink data were collected for two time periods with six months in between. Financial data of two different years were collected to find out which one correlated better with inlink data. Statistically significant correlations were found between inlink data and several financial variables. A comparison between Asian banks and other banks showed that the former attracted significantly more inlinks. This probably reflects the fact that Asian banks overall survived the recent world recession relatively better. The MDS maps generated from co-link data suggest that geographic and linguistic factors determine competitive clusters in the international banking industry. A comparison of the MDS maps from the two different time periods revealed important business information, notably that the Chinese banks moved closer to the major banks from the U.S. and U.K. This is in line with the state of the Chinese banks which emerged from the financial crisis into a stronger position.

Keywords: Competitive intelligence; Web data mining; co-link analysis; Webometrics; financial position; financial performance; banking industry.

1. Introduction

Much research has been carried out in the application of Webometric techniques to commercial websites. Competitive Intelligence (CI) is one of the areas where this type of research has been more successful and promising [1, 2]. According to Kahaner [3], CI consists of a systematic plan to obtain and analyse information about competitors and general trends in the industry. The abundance of digital information available on the Internet has created new opportunities and challenges for companies, therefore they need to monitor changes around them in order to compete in better conditions. The purpose of this research is to analyse the international banking industry by using two different Webometric techniques, inlink analysis and co-link analysis. Inlink and co-inlink are basic Webometric concepts to understand the nature of Web hyperlinks. An inlink, also called back link, is a link pointing to a Webpage, e.g. page X has an inlink coming from page Z. If page X and page Y both have inlinks from page Z, then page X and page Y are co-inlinked [4]. Co-inlink analysis is referred simply as co-link analysis later in this paper. These two Webometric techniques have been proven to be effective in gathering business information in addition to traditional sources. Data gathered through links pointing to Webpages of banks are expected to be a useful additional source of information to generate business knowledge, especially in order to study business competition within the industry. The banking industry has been selected for the study due to its economic impact and degree of internationalization, especially in the context of the recent financial and economic crisis.

Inlink analysis [5] is based on the number of web pages that link to a collection of pages or sites in order to assess their impact. Previous research by Vaughan and colleagues used this technique to explore quantitative relationships between inlink counts to commercial Websites and business performance variables. If a significant correlation exists between these variables, then inlink counts could be used as an indicator of business performance. Vaughan and Wu [6] tested the hypothesis in two groups of Chinese companies: one homogeneous group (in terms of industry) made of China's top 100 Information Technology (IT), and one heterogeneous group made of top 100 privately owned companies. Spearman correlation tests showed significant relationships between inlink counts and three accounting variables: gross revenue, profit, and research and development expenses. Vaughan [7, 8] reinforced this evidence by studying IT industry in China, U.S.A. and Canada. Romero Frías, Vaughan and Rodríguez Ariza [9] extended the research to other industries and found evidence of significant correlations between inlink counts and financial variables (total assets, revenue and net income) in the U.S. banking industry.

Co-link research is based upon the number of webpages that link at the same time two webpages or sites belonging to the set of entities under study. Co-links are analogous to the bibliometric concept of co-citation [10]. The number of co-links to the Websites of a pair of companies is a measure of the similarity between two companies. This means that the more co-links the two

companies have, the more related they are in the views of the sites that link to them. When applied to a set of companies belonging to the same industry, similar companies are understood as competing businesses. Although inlinks among academic websites have been proven to be a relevant measure of similarity [11], competitors very rarely link to each other in order to avoid diverting Web traffic to rival companies [12, 13]. This fact leads to the use of co-links to research competitive positions in an industry.

Vaughan and You [14] applied co-link analysis to map successfully the business competitive positions of 32 telecommunication companies. A more recent paper [15] proposed a method that combines page content (keyword) and Web structure (co-link data) to achieve a more detailed picture of competition in a particular sector (WiMAX) within the telecommunication industry. The methodology developed in these papers has also been tested and verified in other countries and industries, e.g. China's chemical industry and electronics industry [16]. Recently, Romero-Frías and Vaughan [17] extended the use of co-link analysis into the banking industry in the U.S. in order to test the feasibility of combining page content with co-link data to monitor financial crisis.

This study analyses the international banking industry through various Webometric techniques. First of all, we wanted to find out if there is any correlation between the number of Web hyperlinks that a bank attracts and the bank's financial variables. Second, we used co-link analysis to map these banks' global market positions. Both inlink and co-link data were collected in two time periods, December 2008 and June 2009, to find out if and how results changed during the recent world financial crisis. Finally, we compared the Asian banks with other banks to determine if and how they differed in hyperlink patterns. This comparison was guided by the premise of cultural difference and the rapid development of Asian economy.

2. Methodology

2.1. Selection of companies to study

Top 50 banks of the world in terms of total assets (consolidated figures) were selected for the study. The list of companies was gathered on December 4, 2008 from http://www.bankersalmanac.com/addcon/infobank/wldrank.aspx. These banks come from 15 different countries as shown in Table 1. The complete list of all banks, including URLs, labels used in the analysis and inlink counts, is shown in Appendix 1.

Table 1. Country Distribution of the Banks

Country	Australia	Belgium	Canada	China	Denmark	France	Germany	Italy
Number of banks	1	2	2	4	1	6	8	2
Country	Japan	Netherlands	Spain	Sweden	Switzerland	U.K.	U.S.A.	
Number of banks	5	3	2	1	2	6	5	

The Website addresses of each of these banks were collected using Google and then manually checked to ensure its correctness. The vast majority of companies in the study had only one URL for their Websites. For the few that have alternative URLs in the form of alias or redirect, we checked each URL to find out which one had more inlinks and used that one for collecting inlink data. We considered including both URLs in data collection. However, at the time of the study, Yahoo! (the search engine used for data collection) couldn't handle the complex query syntax for collecting co-link data using two URLs.

2.2. Collecting financial data

Financial data of the banks were collected in November 2009, using Mergent database (http://www.mergent.com), a reliable source of business information on global publicly listed companies. The latest financial data available were for the year 2008. The same set of financial data for the year 2007 were also collected in order to compare the correlations to inlink counts. We considered the following three groups of financial variables:

1. Financial position variables: Total Assets.

2. Financial performance variables, in absolute terms: Total Revenue and Net Income.

3. Financial performance variables, in relative terms: Return on Assets (ROA).

Financial data for some banks were not available in the Mergent database, therefore we omitted them in the correlations tests. As shown in table 4, the number of banks in each test varied between 38 to 41. Inlink data were available for all banks in the study, therefore all 50 world banks were included for the co-link analysis.

2.3. Collecting Web link data

Of the three major search engines, Google, Yahoo! and MSN Live Search, only Yahoo! could be

used for data collection for the study. Google's inlink search only returns a sample of all inlinks that the Google database records [18]. Another problem is that Google cannot filter out internal inlinks (inlinks originated from within the Website itself such as "back to home" type of links) as the query term 'link" cannot be combined with any other query terms [19]. In other words, it cannot report the external inlink counts that the study needed. MSN Live Search used to have inlink search functions but the service was turned off around March 2007 [20]. At the time of data collection, December 2008 and June 2009, Yahoo! is the only option for collecting inlink data as required by the study.

Because search engines of different countries may have databases that favour Websites of the host countries [21], we considered using the versions of Yahoo! of the country to which the banks belong. This is feasible for countries such as China as the Chinese version of Yahoo! (www.yahoo.cn) has a database that is different from the global Yahoo! (www.yahoo.com). However, for other countries such as Spain and France, this is not the case. Tests of these versions of Yahoo! showed that they returned the same inlink search results as that from the global version of Yahoo!. This means that the Spanish and French version of Yahoo! just had a different interface but the same underlying database as the global Yahoo!. So the global version of Yahoo! (www.yahoo.com) was used for all data collection.

Yahoo! has two inlink search query terms, link and linkdomain. The "link" query term finds links to a particular page (e.g. link:http://www.abc.com finds links to the homepage of www.abc.com) while the linkdomain query term retrieves all links that point to all pages of a particular Website or domain including the homepage. We used the linkdomain query term for data collection because all links, not just links to homepage, are of relevance to the study. The query syntax for the data collection is illustrated in Table 2 using the hypothetical URLs of www.abc.com and www.xyz.com. We truncated the www portion of the URLs in the queries to capture all links to all subdomains such as mail.abc.com. The "-site:abc.com" part of the query is to filter out internal links coming from within the domain of abc.com itself. Since co-links involve a pair of Websites, the co-link data were collected in the form of a matrix with row x and column y of the matrix representing the number of co-links between URL x and URL y.

Table 2. Illustration of Yahoo! Queries for Data Collection

Types of links searched for	Query
Inlinks to	linkdomain:abc.com –site:abc.com
Co-links between www.abc.com and www.xyz.com	(linkdomain:abc.com –site:abc.com) (linkdomain:xyz.com –site:xyz.com)

We collected first round of Web hyperlink data on December, 2008. At that time, the financial crisis of the world banking industry and the worldwide economic recession was at its peak. We analyzed the data and found some interesting findings as reported in the *Results* section. We then collected the same type of Web hyperlink data again in June 2009 when the world economy began to recover. We wanted to find out if the change of the financial situation would be reflected through the Web hyperlink data.

2.4. Methods of data analysis

Co-link matrix of each round of data was analyzed using multidimensional scaling (MDS) to generate a MDS map. MDS uses a heuristic method to place banks with higher co-link counts closer in the resulting MDS map. The logic of our analysis is that on a macro level, co-links are created for a reason and thus have a pattern. Banks that have similar or related business/services are more likely to be co-linked. In other words, the number of co-links between a pair of banks could potentially be a measure of their similarity or relatedness. The more co-links, the more similar or related would be the business/services of the two banks. Since similar or related banks will be placed closer in the MDS map and banks with similar or related business/services are competitors (two banks offer totally different services are not competing with each other), the MDS map would group competing banks together. We hoped to use MDS the map to see clusters of banks that reveal their market positions. We also hoped to examine of the effectiveness of our methods by comparing the MDS maps from different time periods. If the MDS map would show business relationships among banks, maps from different time periods could potentially reflect changes in business situations.

The raw co-link count collected from Yahoo! were normalized by Jaccard index to obtain a relative measure of the relatedness of the banks. The normalized co-link matrices were feed into SPSS for MDS analysis. The stress values of MDS analysis are 0.075 for the first round of data and 0.073 for the second round of data. These stress values are all fairly low, which suggests a good fit between the data and the MDS map positions.

3. Results

3.1. Correlation between Web inlink and financial data

Before carrying out the correlation tests, we first performed descriptive statistical analysis for both rounds of inlink data. For the December 2008 data, the mean and the median are 133,396 and 52,100 respectively. For the June 2009 data, they are 131,564 and 42,700 respectively. The numbers did not change much over the six month of time period, suggesting that inlink counts are

a fairly stable over time. However, if we analyze the changes at the individual bank level, we find that inlink counts for some companies have changed significantly, as shown in Table 3.

Table 3. Top changes in inlink counts from December 2008 to June 2009

	Top 5 banks with increased inlink count		Top 5 banks with decreased inlink count	
1	China Construction Bank Corporation	77.30%	The Norinchukin Bank	-54.06%
2	DZ Bank AG	48.19%	Wachovia Bank NA	-41.33%
3	Kreditanstalt für Wiederaufbau (KfW)	41.53%	Calyon	-39.54%
4	Agricultural Bank of China	41.45%	Fortis Bank SA/NV	-39.12%
5	Deutsche Bank AG	33.28%	Dresdner Bank Group	-39.02%

German banks were among the top with increased inlink counts. This could be explained by some economic facts. For instance, KfW (41.53% more inlinks) was back to earning profits in the first quarter of 2009 after two years of heavy losses resulting from the crisis of financial markets. Also, the interim accounts of the KfW Group at 31 March 2009 closed with a consolidated profit of EUR 80 million. Finally, Deutsche Bank, the biggest German bank, had a 33.2% increase in inlink count. It reported a net income of EUR 1.2 billion for the first quarter 2009 compared to a net loss of EUR 141 million for the same period of 2008. Also, in the first quarter of 2009, Deutsche Bank accelerated its plan in taking over the country's biggest retail bank, Deutsche Postbank [22]. Among the top five banks with increased inlinks, two are from China. It is known that the Chinese banks overall weathered the financial crisis better because of the tighter government regulations [23].

In contrast, banks, such as Wachovia, Fortis and Dresdner, had a decreased inlink counts for the same period. This situation could be interpreted as investors and other stakeholders losing interest in these companies. Again some economic information could help to understand the reasons. Wachovia Corporation (41.33% decrease of inlink count) was purchased by Wells Fargo on December 2008, and consequently it ceased to be an independent corporation on that date. The Wachovia brand is expected to be absorbed into the Wells Fargo brand over the next three years. Fortis (39.12% decrease of inlink count) underwent serious financial problems over the last two years and most of the company was sold in parts in 2008. Finally, Commerzbank announced that it would acquire Dresdner Bank (39.02% decrease in inlink count) on August 31, 2008. This EUR 9 billion takeover was finalized in the first quarter of 2009 [22]. Although, this move will create a

second major of national banking champion together with Deutsche Bank, the decrease in the number of inlinks to Dresdner Bank Website suggests a loss of attention in the acquired entity. As indicated on the Website of Dresdner Bank, it is a brand of Commerzbank since May 2009. The situation of Norinchukin Bank could be understood by the general bad situation of the Japanese banking sector, especially by contrast to the strength of other Asian banks, i.e. the Chinese banks [24].

The frequency distribution of the two sets of inlink counts (December 2008 and June 2009) are very skewed, so the Spearman correlation test rather than the Pearson correlation test was used. The correlation coefficients between the inlink counts and the financial performance data are shown in Table 4. All correlation coefficients are statistically significant except for the variable return on assets.

Table 4. Correlation between inlink data and financial data

Spearman's rho	Financial year	N	Inlinks of December 2008	Inlinks of June 2009
Total Assets	2008	39	.58**	.54**
	2007	39	.44**	.43**
Total Revenue	2008	38	.62**	.60**
	2007	39	.44**	.40*
Net Income	2008	39	.56**	.52**
	2007	39	.57**	.55**
Return on assets (Net)	2008	39	.23	.19
	2007	39	.02	.06
Number of employees	2008	41	.80**	.80**
	2007	39	No employee data for 2007	No employee data for 2007
**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).				

Findings show that inlink counts could be use as an indicator of the banks' financial position and financial performance measures. However, there is no significant relationship between inlink count and return on assets, a relative financial performance measure. This makes sense since the

number of inlinks retrieved from the commercial search engine includes all the links pointing to a particular bank Website since its creation. This means that the inlink count is accumulative in nature. The same could be said about the financial variables that are measured in absolute terms, especially for the total assets variable. It is expected that a company becomes bigger over time and therefore its assets increase. Although performance variables such as total revenue and net income are calculated for a particular period of time, they usually increase over the years, e.g. total revenue is expected to be higher in the fifth year of a company than in its first year. By contrast, return on assets is not dependent of the company's size and therefore it is logical that no significant correlation is found. It would be interesting to analyse the changes of inlink counts over several years to calculate a relative measure for this variable.

The correlation is stable over time as suggested by the very close match between the two sets of correlation coefficients. The higher correlations are found between the Dec. 2008 inlink data and the 2008 financial data, i.e. when the year of inlink data matches that of the financial data. If we compare the correlation coefficients between inlink counts and financial data of 2007 and 2008, we observe that the relationship is stronger for the 2008 financial data. The only exception is net income which presents very similar correlation coefficient in both cases although slightly higher for the 2007 financial data. This indicates that matching the observation period could be an important issue in order to obtain a more accurate predication of one variable based on the other. It also suggests that inlink data are somewhat time sensitive, which is a desirable feature.

These correlation coefficients are consistent with the coefficients reported in Romero Frías, Vaughan and Rodríguez Ariza [9] for the U.S. banking industry as shown in Table 5.

Table 5. Correlation between inlink data and financial data for the U.S. banking industry [9]

	Total Assets	Total Liabilities	Total Revenue	Net Income	Return on Assets
January 2009 inlink data	0.74**	0.73**	0.75**	0.63**	0.13
May 2009 inlink data	0.70**	0.69**	0.71*	0.62**	0.18
**. Correlation is significant at the 0.01 level (2-tailed). *. Correlation is significant at the 0.05 level (2-tailed).					

Correlation coefficients for U.S. banking industry are higher than that for the global banking industry. This is explained by the homogeneous competitive conditions that exist in the U.S. market compared with the heterogeneous markets where the banks included in this study operate. Different countries have different economical and financial conditions, which could explain the

lower correlations when banks from many countries are analyzed together. The extent to which Internet is used for commercial purposes in different countries could also be a factor.

It is important to interpret the correlations found in the study appropriately. Proving that a correlation is statistically significant does not prove causation. The large number of inlinks that a bank's Website attracted does not cause the better financial performance, although a positive Web image as suggested by the larger number of inlinks may contribute positively to the bank's business. A likely explanation of the correlation is that a bank that is doing well would have better financial data and also being able to maintain a high profile on the Web attracting larger number of inlinks. In other words, the larger number of inlinks may be a symptom rather than a direct cause of the good economic performance of a bank. Although we cannot establish a causal relationship, the correlation found is still very useful information. Now that we know that the two variables are correlated, we can predict one based on the other. Because the Web inlink data are publically accessible and can be collected easily, we could find trends of a bank's financial performance in the long run based on inlink data. The fact that we were not able to locate financial data for some banks underscores the usefulness of this approach. For non listed companies, financial data are most likely unavailable. The correlation that we found is potentially more useful in situation of scarce or no financial data available. Another use of the correlation found in the study is to identify companies whose Web presence is not on par with their financial performance. The correlation could be used to develop a regression equation and we could use the equation to predict what a company's Web position should be based on their financial data. Comparing the predicted Web position with the real one would reveal those that are underperforming on the Web. Further analysis could then be done to find out why and how to improve.

3.2. Inlink count comparison between Asian banks and other banks

Due to the particular features of the Asian economy and the specific positions of the Chinese and Japanese banks revealed by our co-link analysis (reported later in section 3.3), we decided to determine if there is a statistically significant difference in inlink counts between Asian banks and other banks in the study. A Mann-Whitney test was carried out for this purpose. The tests indicate that inlink counts are significantly different ($p=0.042$ for Dec. 2008 data and $p=0.024$ for June 2009 data). As shown in Table 6, Websites of the Asian banks attracted more inlinks than their counterparts elsewhere. This is in the context that the search engine used for data collection, the global Yahoo!, may only over-represent the U.S. Websites [25], so the higher number of inlinks to Asian bank sites is likely to be very strong evidence. Note that the inlink count discrepancy between Asian banks and other banks increased from Dec. 2008 to June 2009. These are important findings because they reflected the financial power of Asian banks, in particular the Chinese banks, as demonstrated during the recent world financial crisis. The Mergent report of the

Banking industry in the Asia Pacific region [24] underlines the good performance of Chinese banks. Some factors contributing to this are: banking reforms, rises in foreign investment, stronger supervision and more intense competition. At this moment, China has emerged as a prominent player on the international stage and its banks remain stronger and healthier compared with many companies in the U.S. and Europe. This strength compensates the weakness of the Japanese banking industry, which is also included in the Asian group for the analysis.

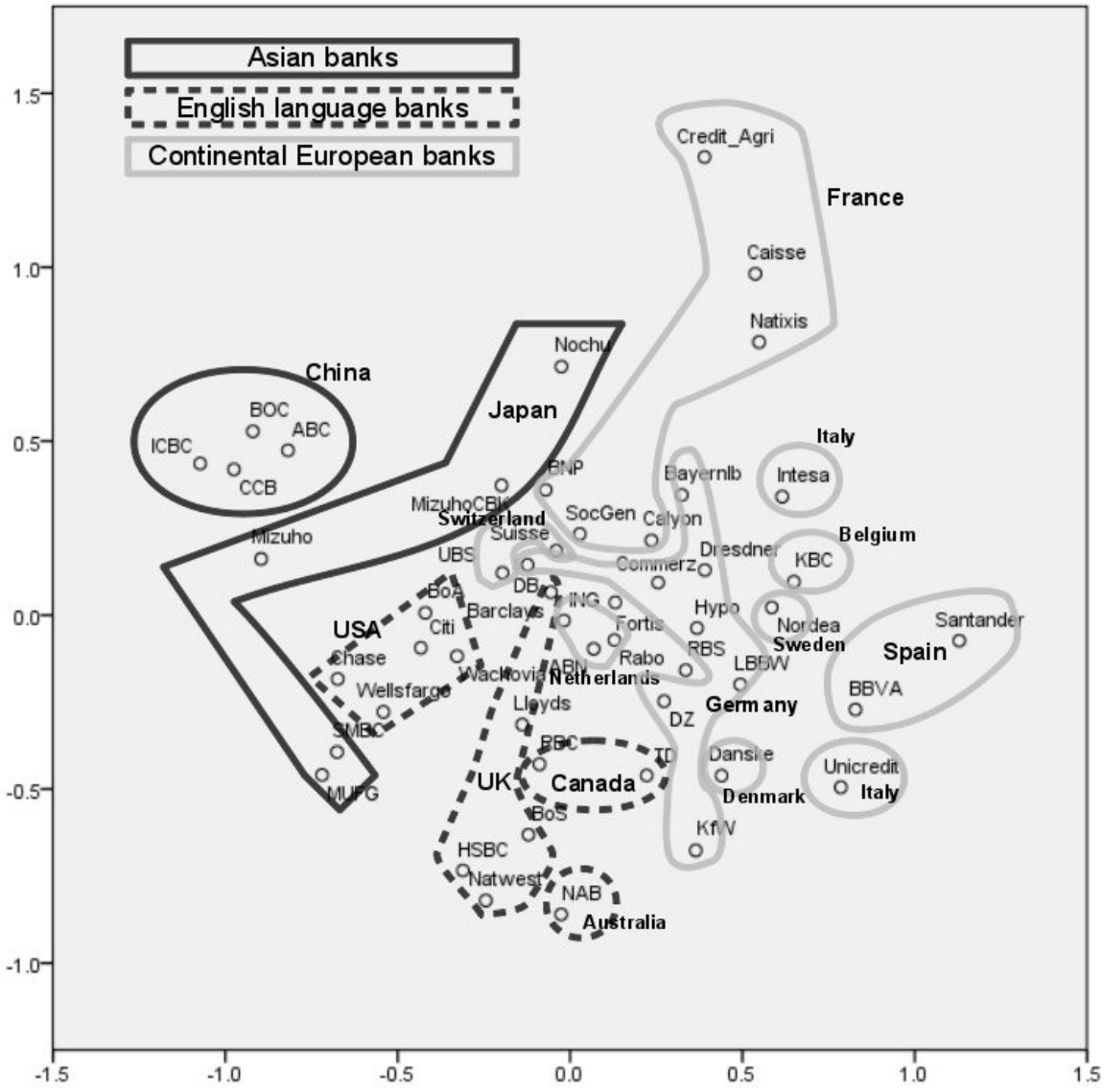
Table 6. Inlink count comparison between Asian Banks and other banks

	Number of banks	Dec 2008 median inlink count	June 2009 median inlink count
Asian Banks	9	193,000	273,000
Other banks	41	40,600	35,900

3.3. Co-link analysis

MDS maps (see Figure 1 and 2) based on the number of co-links between banks show the relative competitive positions of the banks in the study. The MDS map based on Dec. 2008 data is shown in Figure 1. It presents a clear pattern that reflects national, language and regional clusters. Three main areas are identified: Asian banks, English language banks and European banks (excluding U.K.). Chinese and Japanese banks are clustered together in one side of the map. Chinese banks are located at the outskirt of the map isolated from other companies. This indicates that they compete mainly at local level. They receive many inlinks but relatively few of them are co-links with other banks. Japanese banks are in an intermediate position with MUFG and SMBC close to U.S. banks, Mizuho to Chinese banks and Nochu and MizuhoCBK to European entities. Banks from English speaking countries (U.S., U.K., Canada and Australia) are clustered together, although national groups are also clearly identified within the cluster. Continental European banks (from Belgium, Denmark, France, Germany, Italy, Netherlands, Spain, Sweden and Switzerland) occupy the rest of the map. The two Swiss banks are in a central position among all banks studied, mirroring the central and competitive position of the Swiss banks. Dutch banks and some other banks that are not clustered with their national banks (Fortis and Royal Bank of Scotland) are also in a central position. The cluster for the French banks is clearly identifiable, likewise the German group although it is a bit more dispersed.

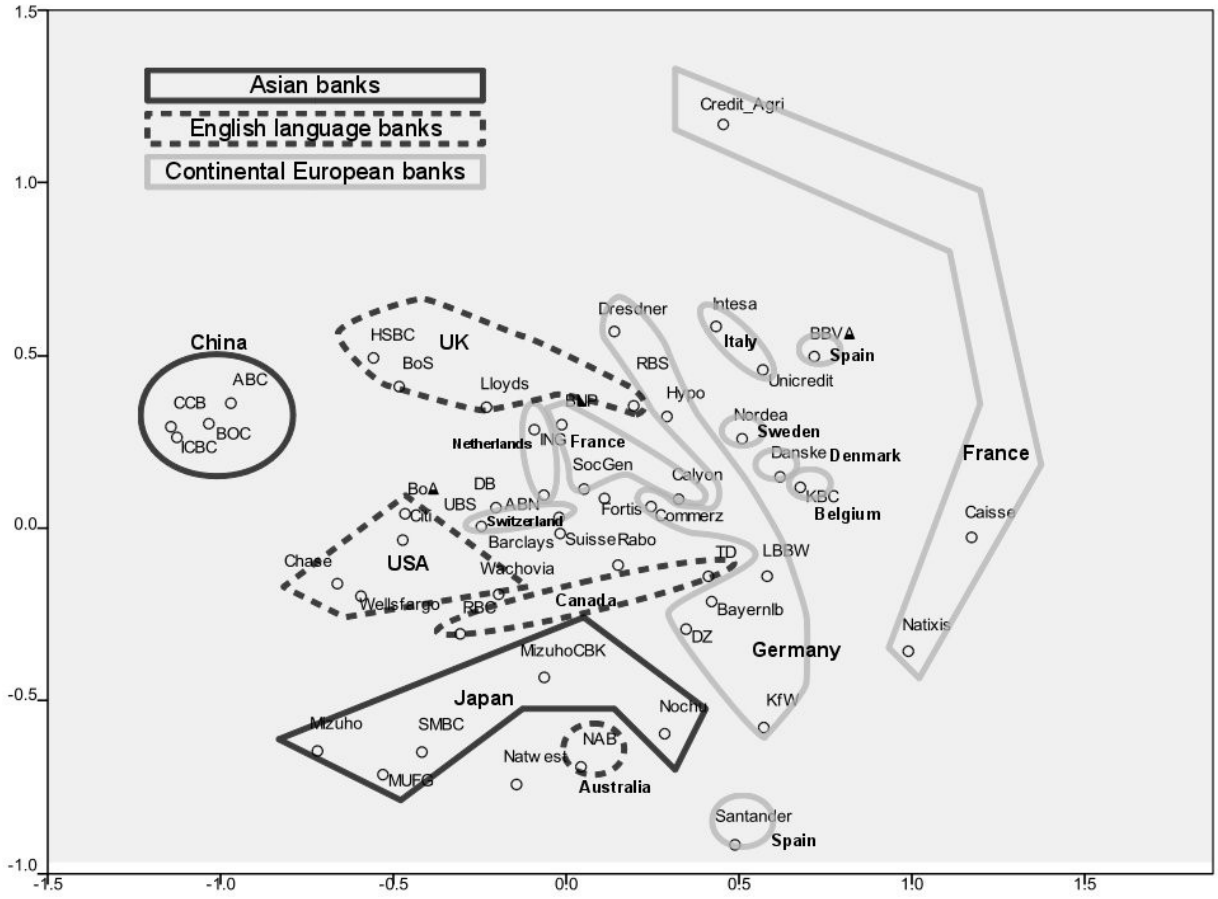
Figure 1. MDS map for based on December 2008 data



The MDS map based on June 2009 data is shown in Figure 2. To make a clear comparison with Fig. 1, banks in Fig. 2 are also clustered and labelled based on country. Despite the fact that the country factor remains salient, there are some significant changes from Fig. 1 to Fig. 2 that are likely to be related to the evolution in the economic and financial crisis over the six month period. It is worth noting that the Chinese and Japanese banks moved apart and the U.S. banks are placed between them. The Chinese banks also moved closer to the British banks compared with their position in Fig. 1. Although further monitoring and analysis is needed to find out for sure whether this reflects the changing perception the world has on the Chinese banks as the result of the world financial crisis, it is reasonable to speculate that this reflects the fact that major Asia-Pacific banks, except the Japanese banks, remain strong and well-positioned in the international context [24]. This is particularly true of the Chinese banks that are among the top in the world. Swiss banks and

major European banks (Deutsche Bank and Barclays) maintain their central positions. This reflects the strong and stable positions these banks have in the industry. French banks are now divided into two groups.

Figure 2. MDS map based on June 2009 data



4. Conclusions and future research

The study achieved its goal of using a combination of inlink and co-link analysis to provide a general analysis of the global banking industry. We found significant correlations between inlink counts and various financial variables, some of them have not been tested in previous studies. Findings are consistent with previous research in the same industry for a particular country [9]. This confirms that Web hyperlink data can provide useful business information even in a heterogeneous environment where companies from multiple countries are included. A comparison of the inlink count between Asian banks and other banks showed that Asian banks received more inlinks. This is consistent with the fact that Asian banks weathered the recent world financial crisis relatively better. This finding also further shows the usefulness of Web hyperlink data in gathering business information. MDS analysis based on co-links indicates that the global banking industry is

partially organised in regional markets despite the existence of main global players due to the internationalization process of financial activities. Some of them seem to be more isolated such as the Chinese banks for the first time period of the study. This is possibly one of the reasons why the Chinese banks didn't suffer from the recent world financial crisis as much as other banks did. However, data collected from the second time period revealed changing positions of the Chinese banks. They moved closer to major players such as U.S. and U.K. banks. This probably reflected the changing perception of the global financial industry toward the Chinese banks as the result of their relative resilience in the world financial crisis. Swiss banks are located at the center of the map together with major banks (e.g. Barclays or Deutsche Bank) from different European countries. This echoes the central role these banks play in the world banking industry.

There could be some limitations in the inlink and co-link analysis due to the inherent limitations of the search engine used for data collection. Yahoo! maintains a different database in some countries, e.g. China. Although this did not prove to be a problem for this particular study as the inlink counts for Asian countries were actually higher, it could be potential problem in other studies. The number of banks from each country is not homogeneous and this could affect the results as the main variable used in the analysis is the bank's country origin. On the other hand, if we select banks with similar size or scale, the number of banks from each country will always be uneven because different countries have different economical power, so the choice must be made that takes various factors into consideration. It is also worth to note that the Web policy of the banks is a significant issue.

Future research will focus on monitoring the changes of the MDS maps by collecting and analyzing data over time. We could also expend the scale of the study to include more banks from more countries to find out if the findings from this study can be generalized. Although Webometrics is mainly quantitative nature, it could be complemented by qualitative analysis. We could do a content analysis of the linking pages to further understanding the Web linking phenomenon. . Finally, it will be useful to consolidate findings from related studies to develop a systematic methodology of Webometrics for competitive intelligence.

5. References

[1] B. Tan, S. Foo and S.C. Hui, Web information monitoring for competitive intelligence, *Cybernetics and Systems*, 33(3) (2002) 225-251.

[2] E. Reid, Using Web link analysis to detect and analyze hidden Web communities. In: D. Vriens (ed.), *Information and Communications Technology for Competitive Intelligence*, (Ideal Group Inc, Ohio, 2003) 57-84.

[3] L. Kahaner, *Competitive Intelligence – How to Gather, Analyze, and Use Information to*

Move Your Business to the Top (7th ed., Touchstone, New York, 1996).

[4] E. Garfield, *Citation indexing: Its theory and applications in science, technology and the humanities* (Wiley, New York, 1979).

[5] M. Thelwall, *Link Analysis: An Information Science Approach* (Academic Press, San Diego, 2004).

[6] L. Vaughan and G. Z. Wu, Links to commercial websites as a source of business information, *Scientometrics*, 60(3) (2004) 487–496.

[7] L. Vaughan, Exploring website features for business information, *Scientometrics*, 61(3) (2004) 467-477.

[8] L. Vaughan, Web hyperlinks reflect business performance—A study of US and Chinese IT companies, *Canadian Journal of Information and Library Science*, 28(1) (2004) 17–31.

[9] E. Romero Frías, L. Vaughan and L. Rodríguez Ariza, El recuento de enlaces a sitios Web comerciales como indicador de las variables de desempeño y posición financiera de la empresa: estudio empírico de diversos sectores empresariales en Estados Unidos (Unpublished proceedings XV Congreso de la Asociación Española de Contabilidad y Administración de Empresas, Valladolid, September 23-25, 2009).

[10] H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents, *Journal of the American Society for Information Science*, July-August (1973) 265–269.

[11] M. Thelwall and D. Wilkinson, Finding similar academic Web sites with links, bibliometric couplings and colinks, *Information Processing and Management*, 40 (2004) 94–101.

[12] D. Shaw, Playing the links: interactivity and stickiness in .com and “not.com” websites, *First Monday*, 6(3) (2001). Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/837/746 (accessed 10 May 2009).

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

[14] L. Vaughan and J. You, Comparing business competition positions based on Web co-link data—The global market vs. the Chinese market, *Scientometrics*, 68(3) (2006) 611–628.

[15] L. Vaughan and J. You, Content assisted web co-link analysis for competitive intelligence, *Scientometrics*, 77(3) (2008) 433-444.

[16] L. Vaughan, J. Tang and J. Du, Examining the robustness of Web co-link analysis, *Online Information Review*, 33(5) (2009) 956-972.

[17] E. Romero-Frías and L. Vaughan, Financial distress of U.S. banking industry viewed through Web data. In *Proceedings of ISSI 2009 – the 12th International Conference of the International Society for Scientometrics and Informetrics* (Rio de Janeiro, Brazil, July 14-17, 2009) 206-210.

[18] Google, *Links to your site (2009)*. Available at: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=55281 (accessed 3 June 2009).

[19] Google, *Google SOAP Search API Reference (2006)*. Available at: http://www.google.com/apis/reference.html#2_2 (accessed 3 June 2009).

[20] Live Search, *We are flattered, but... (2007).* Available at: http://www.bing.com/community/blogs/search/archive/2007/03/28/we-are-flattered-but.aspx (accessed 3 June 2009).

[21] L. Vaughan and M. Thelwall, Search engine coverage bias: Evidence and possible causes, *Information Processing & Management*, 40(4) (2004) 693-707.

[22] Mergent, Europe - Banking Sectors (September 2009), available at http://webreports.mergent.com (accessed 24 December 2009).

[23] J. Shaw and G. Parussini, *China bank regulator: Global regulatory cooperation very fragile (2009)*. Available at: http://www.nasdaq.com/aspx/stock-market-news-story.aspx?storyid=200910051100dowjonesdjonline000261&title=china-bank-regulatorglobal-regulatory-cooperation-very-fragile (accessed 15 October 2009).

[24] Mergent, Asia-Pacific - Banking Sectors (June 2009), available at http://webreports.mergent.com (accessed 24 December 2009).

[25] L. Vaughan and Y. Zhang, Equal representation by search engines? A comparison of Websites across countries and domains, *Journal of Computer-Mediated Communication*, 12(3) (2007), available at http://jcmc.indiana.edu:80/vol12/issue3/vaughan.html

Appendix 1. Banks included in the study

Labels	Company	Main URL	Country	Dec 2008 inlink data	June 2009 inlink data
RBS	The Royal Bank of Scotland Group plc	http://www.rbs.com	U.K.	26,700	35,200
DB	Deutsche Bank AG	http://www.db.com	Germany	66,100	88,100
BNP	BNP Paribas SA	http://www.bnpparibas.com	France	143,000	112,000
Barclays	Barclays PLC	http://www.barclays.com	U.K.	42,900	28,200
Credit-Agri	Crédit Agricole SA	http://www.credit-agricole.com	France	28,700	30,900
UBS	UBS AG	http://www.ubs.com	Switzerland	147,000	148,000
SocGen	Société Générale	http://www.socgen.com	France	54,300	42,000
ABN	ABN AMRO Holding NV	http://www.abnamro.com	Netherlands	52,100	42,700
Unicredit	UniCredit SpA	http://www.unicreditgroup.eu	Italy	79,900	50,000
ING	ING Bank NV	http://www.ing.com	Netherlands	58,800	64,900
MUFG	The Bank of Tokyo-Mitsubishi UFJ Ltd	http://www.mufg.jp	Japan	364,000	454,000
Santander	Banco Santander SA	http://www.santander.com	Spain	395,000	364,000
Chase	JPMorgan Chase Bank National Association	http://www.chase.com	U.S.A.	223,000	144,000
BoA	Bank of America NA	http://www.bankofamerica.com	U.S.A.	339,000	304,000
Citi	Citibank NA	http://www.citibank.com	U.S.A.	140,000	98,700
Suisse	Credit Suisse Group	http://www.credit-suisse.com	Switzerland	61,500	70,100
Fortis	Fortis Bank SA/NV	http://www.fortis.com	Belgium	34,000	20,700
ICBC	Industrial & Commercial Bank of China Limited	http://www.icbc.com.cn	China	985,000	937,000
CCB	China Construction	http://www.ccb.com	China	304,000	539,000

	Bank Corporation				
BoS	Bank of Scotland plc	http://www.bankofscotland.co.uk	U.K.	13,700	9,390
HSBC	HSBC Bank plc	http://www.hsbc.co.uk	U.K.	85,100	65,700
Intesa	Intesa Sanpaolo SpA	http://www.intesasanpaolo.com	Italy	47,800	46,700
SMBC	Sumitomo Mitsui Banking Corp.	http://www.smbc.co.jp	Japan	105,000	103,000
Commerz	Commerzbank AG	http://www.commerzbank.com	Germany	10,600	9,100
Calyon	Calyon	http://www.calyon.com	France	10,900	6,590
Rabo	Rabobank Nederland	http://www.rabobank.com	Netherlands	11,100	11,300
Dresdner	Dresdner Bank Group	http://www.dresdner-bank.com	Germany	5,280	3,220
Caisse	Caisse Nationale des Caisses d'Epargne et de Prévoyance	http://www.caisse-epargne.com	France	28,800	31,700
Lloyds	Lloyds TSB Group plc	http://www.lloydstsb.com	U.K.	40,400	28,900
ABC	Agricultural Bank of China	http://www.abchina.com	China	193,000	273,000
BOC	Bank of China Limited	http://www.boc.cn	China	399,000	485,000
Danske	Danske Bank A/S	http://www.danskebank.com	Denmark	7,990	7,820
Wachovia	Wachovia Bank NA	http://www.wachovia.com	U.S.A.	143,000	83,900
RBC	Royal Bank of Canada	http://www.royalbank.com	Canada	40,600	40,500
Hypo	Bayerische Hypo- und Vereinsbank AG	http://www.hypovereinsbank.de	Germany	14,000	16,300
Natixis	Natixis	http://www.natixis.fr	France	7,690	4,720
Nochu	The Norinchukin Bank	http://www.nochubank.or.jp	Japan	5,790	2,660
DZ	DZ Bank AG	http://www.dzbank.de	Germany	9,650	14,300

Natwest	National Westminster Bank Plc	http://www.natwest.com	U.K.	21,800	18,300
Nordea	Nordea Group	http://www.nordea.com	Sweden	41,200	35,900
Mizuho	Mizuho Bank Ltd	http://www.mizuhobank.co.jp	Japan	124,000	113,000
LBBW	Landesbank Baden-Württemberg	http://www.lbbw.de	Germany	6,190	7,070
MizuhoCBK	Mizuho Corporate Bank Ltd	http://www.mizuhocbk.co.jp	Japan	5,710	3,570
KfW	Kreditanstalt für Wiederaufbau	http://www.kfw.de	Germany	44,300	62,700
BBVA	Banco Bilbao Vizcaya Argentaria SA	http://www.bbva.com	Spain	43,100	53,200
NAB	National Australia Bank Ltd	http://www.nab.com.au	Australia	19,800	21,200
Wellsfargo	Wells Fargo Bank NA	http://www.wellsfargo.com	U.S.A.	346,000	291,000
Bayernlb	Bayerische Landesbank	http://www.bayernlb.de	Germany	5,810	4,700
KBC	KBC Bank NV	http://www.kbc.com	Belgium	17,000	20,300
TD	The Toronto-Dominion Bank	http://www.td.com	Canada	37,300	39,800

4 DISCUSSION

4.1 Background

During its 20 years of life, the Web has become a laboratory for social sciences. Since its creation, scientists have struggled to study and analyze the diverse phenomena that takes place in the Web, which is characterized by the large amounts of data available and its continuous transformation and growth. The massive collection of information makes it possible to describe the Web as an enormous unstructured and heterogeneous database that, despite its appearance, is not randomly built. Therefore, Web data can be exploited from different perspectives (based on content, structure and user behavior) in order to study the unique online phenomena or offline phenomena reflected in the Web.

The combination of science and the Internet is known by various terms, such as e-Science or e-Research. e-Science refers to large scale science that is carried out through distributed global collaborations enabled by the Internet. Many e-Science initiatives have underlined the importance of distributed computational power and grid computing, although most of the time, especially when referred to Social Science, the collaboration between researchers through the Internet does not require this type of resources. For instance, the investigation based on information about the content and structure of the Web does not necessarily require large computational resources. Many databases are available online, for example, search engines could be used to collect data in order to carry out research on different aspects of the Web. Search engines crawl large parts of the Web and provide information about the content of the Web pages and the links that form the structure of the Web. This data could be used to reveal patterns that otherwise would remain hidden from us. A potential source of information on the Web derives from visualizing hyperlink networks. The exploitation of this type of information has been done by applying data mining techniques to the Web. In this line, Webometrics has emerged as a new discipline that applies to the Web concepts and methods derived from bibliometrics and information science (see chapter 3) .

Webometrics is defined by Björneborn (2004; in Björneborn and Ingwersen, 2004: 1217) as "the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches". However, the study of the Web differs significantly from the study of scientific production and academy. The differences could be understood through the analysis of some of the main features of the Web. The characteristics of the Web could also help us to grasp the implications and limitations of the findings obtained in this type of research. For instance, the dynamic nature of the Web (Ke *et al*., 2005) implies that any research based on Web data is always a snapshot of that particular moment and state of the Web, subject to the limitations of the tools used to gather the information (for

example, search engines) and impossible to be reproduced in future times.

As mentioned, commercial search engines are one of the most important sources of Web data, fundamentally when we seek to analyze the entire Web. However there are some significant limitations derived from the use of commercial search engines. Here we summarize some of those limitations that can be useful to highlight some of the specificities of the Web as a matter of research:

- Search engines do not index the entire Web (Lawrence & Giles, 1998, 1999; Sherman and Price, 2001; Bar-Ilan, 2004; Thelwall, Vaughan and Björneborn, 2005). Sometimes this is not due to technical reasons but due to the design of the Web sites, which could ban Web crawlers to access some of their parts (Koster, 2009a, 2009b). Nevertheless, Vaughan and Zhang (2007) found the percent of coverage to be much higher.
- Ranking systems eliminate similar or identical pages in their results, in order to avoid providing useless information (Gomes and Smith, 2003; Thelwall, 2008).
- Crawling and reporting algorithms are commercial secrets and, therefore the exact criteria used to rank the information are unknown (Thelwall, Vaughan and Björneborn, 2005).
- The total numbers of results offered by search engines are estimates as they use algorithms that prioritize response time rather than exhaustiveness (Björneborn and Ingwersen, 2001).
- Results can be subject to national or language biases (Vaughan and Thelwall, 2004).
- The results can fluctuate and change over the time (Bar-Ilan, 2000; Mettrop & Nieuwenhuysen, 2001). In addition, only a few number of pages are accessible (usually just a maximum of 1000), creating problems when random sampling is needed, for instance. Nevertheless, some research also indicates that search engines have become more stable (e.g. Thelwall, 2001a; 2001b; Vaughan & Thelwall, 2003; Vaughan, 2004a). Uyar (2009) has recent findings about accuracy of search engine hit count.

Even with such limitations, earlier studies in different matters have discovered patterns in linking networks and significant correlations between online and offline phenomenon validating the results through link classifications. Nevertheless, it is essential to keep in mind these limitations when analyzing and interpreting Web data.

This thesis focuses on the analysis of hyperlinks, this is, in the analysis of the structure of the Web. A hyperlink can be defined as a reference or navigational element in a document to another section of the same document or to another document. Somehow, the links constitute the hidden

structure of the Web connecting different sites and Web pages that would stay isolated unless the specific URL is known (Berners-Lee, 1999). They can be regarded as an endorsement of a target page, especially if the creator has placed that link because it points to a useful or relevant resource. This idea resembles clearly the Garfield's proposal (1979) to use bibliographic citations as the basis to rank scientific production, a technique that is still used to evaluate the quality of academic research. The creation of hyperlinks is not an irrelevant phenomenon, but implies significant social repercussions (Turow and Lokman, 2008). Exploitation of hyperlinks is well illustrated by the functioning of commercial search engines (Batelle, 2005); for instance, Google's search engine dominance is derived from the exploitation of *Pagerank* system (Brin and Page, 1998).

The application of Webometric techniques to commercial Web sites is not as developed as in the academic field (Thelwall, Vaughan and Björneborn, 2005). Competitive Intelligence (Kahaner, 1996) is one of the areas where this research has been more successful and promising (Tan *et al*., 2002; Reid, 2003). We believe that the application of Webometric techniques and other Internet research methodologies to companies could provide new resources for companies to generate and maintain competitive advantages (Porter, 1980, 1985). For example, a recent study (Choi and Varian, 2009), carried out by Google researchers, used data from the Google Trends service to anticipate consumer behavior in several industries.

The empirical research developed in this thesis was based on two approaches: link impact analysis and co-link analysis. On one hand, link impact analysis is based on the comparison of the number of Web pages or Web sites that are linked to a set of Web pages or Web sites under research. The purpose of this type of research is, according to Thelwall (2009: 28), "to evaluate whether a given website has a high link-based web impact compared to its peers". Inlink counts can also be an indirect gauge of other attributes of the organization represented by the Web site. For instance, this has been traditionally used, within the academic field, as a potential estimator of research performance (Smith and Thelwall, 2002). Concerning business Web sites, Vaughan (2004a, 2004b) and Vaughan and Wu (2004) found significant evidence of the positive relationships between the number of links received by a business Web site and financial variables. The studies included in the sections 3.1.1 and 3.1.2 extend this research by finding evidence in different countries and industries. The study in section 3.1.3 represents the first attempt to determine the explanatory variables of the number of inlinks received by a business Web site.

On the other hand, co-link analysis could be considered as a type of link relationship mapping technique (Thelwall, 2009). These are based on the link data that interconnect a set of Web sites in different ways in order to draw a diagram that illustrates the relationships between them. In particular, co-link analysis is based upon the number of Web pages that link at the same time two

Web pages or sites belonging to the group of entities under study. Co-links are analogous to the bibliometric concept of co-citation (Small, 1973). Co-link analysis has also been demonstrated to be a useful tool to reveal the cognitive or intellectual structure of a particular field of study (Zuccala, 2006). This method is particularly useful when Web sites interlink each other which happens very rarely. It is the case of commercial Web sites that scarcely link the Web site of a competing company, especially when they are in the same industry (Vaughan, Gao and Kipp, 2006). The explanation to this could be that companies seem to avoid diverting Web traffic to competitors (Shaw, 2001). Moreover, as Vaughan (2006) points out, co-link data are more robust than inlink data as the former are less easily manipulated. Web co-link analysis for business information started by focusing on a single industry (Vaughan and You, 2006) or on a specific sector within an industry (Vaughan and You, 2008; Vaughan and You, 2009). The studies included in the sections 3.2.1 and 3.2.2 explored the use of co-link analysis for investigating companies belonging to different industries and to analyze the evolution of financial crisis in the American banks.

Finally, the research in section 3.3.1 combines both methods to provide an overview of the international banking industry.

4.2 Research questions: findings and contributions

Several aspects of business Web sites were studied in this thesis, some results were expected while others were not. Webometric methods were used to analyze the linking patterns to business Web sites, by extending the evidence found by previous research. As far as we know, this is the first approach to Webometric research from the area of business studies. Therefore one of the main contributions of this thesis is to present and develop new methods of research on business issues by profiting from a interdisciplinary approach.

From a broad perspective, the goal of this research was to establish a transversal link, using the jargon presented in the paper, between business studies and information science studies in the field of Web research. Moreover, extending previous research on this area, we tried to improve the understanding of linking patterns on commercial Web sites. To achieve this goal, we worked in three main lines:

- to verify whether the correlations found between inlink counts and financial variables existed in different industries, regions and types of companies;
- to investigate whether co-link analysis could provide us with knowledge about economies in general; and,
- to combine previous methods to provide an overall approach to a single industry.

The goals of this thesis were summarized in five research questions that we tried to answer throughout different studies.

4.2.1 Research question 1: To what extent financial variables and business web presence, measured by inlink counts, are related?

The first research question was answered by studying the relationships between inlink counts and financial variables in different world regions and in multiple industries. We found that in general terms the inlink data collected correlated well with offline financial variables as shown by previous research. The correlations between the online and the offline data validated the conclusions that were already found. Previous studies (Vaughan, 2004a, 2004b; Vaughan and Wu, 2004) were limited to the Information Technology industry and to three countries: China, USA and Canada. It was necessary to extend the investigation:

- to a wider variety of industries; and
- to other regions, especially Europe, one of the most important markets in the world.

In addition, some companies in the studies were private entities as we included all the companies that have financial data in the database we used. The majority of European companies in the Amadeus database were private.

The study in section 3.1.2 extended previous research by finding evidence to significant correlations in several industries in the United States. The study analyzed five different industries (commercial banks, construction of buildings, general merchandise store, utilities and mining), as well as the companies in the Dow Jones Industrials. The results revealed that there are significant correlations in industries that not only pertain to Information Technology, indicating that hyperlink data could be used as a meaningful variable in multiple industries, not exclusively to information-centered ones.

In section 3.3.1 we analyzed the top 50 international banks. Spearman's test indicated that the majority of the correlation coefficients between inlinks and a set of financial variables (i.e. total assets total revenue, net income and earnings before tax) were significant at the .01 level. Only return on assets (ROA) was not found to be statistically significant. Correlation coefficients for U.S. banking industry are higher than those for the global banking industry. This is explained by the homogeneous competitive conditions that exist in the U.S. market compared to the heterogeneous markets where the banks included in this study operate. Different countries have different economical and financial conditions, which may explain the lower correlations when banks from many numerous countries are analyzed together. The extent of which the Internet is used for commercial purposes in different countries could also be a significant factor.

Finally, the study in section 3.1.1 extends the research into the European context. Spain and the

United Kingdom were selected as they represent two big economies in the European Union and their two languages are among the most commonly used on the Web. The study confirmed, in the European context, findings from the previous studies. It has also been found that the Web hyperlink data reflected some unique features of the EU economy. For example, the correlations between the inlink data and the financial data do not change significantly when data from the two countries are merged, reflecting that two countries share a common market. When comparing results from the current study with that from previous studies of other countries such as the U.S., the study also found issues that need further exploration to gain a deeper understanding of the relationship between Web data and financial variables. The majority of studies so far indicate that only when companies belong to the same industry the correlations are significant. However, there are contradictory evidence as the study in Spain and United Kingdom shows significant correlations when different industries in the same country are put together.

From these studies included in the thesis we could say that the number of links received by a business Web site is strongly correlated with financial position measures (total assets) and, to a lower extent, with financial performance measures (total revenue). Profit (and loss) and some relative financial performance ratios, such as Return on Equity (ROE) and Return on Assets (ROA), are rarely found significant. The findings prove that inlink counts could be used as an indicator of the business financial position and financial performance measures in absolute terms. However, there is no evidence to this theory when we refer to relative financial measures, such as ROA. This could be explained by the nature of the variable “inlink” that represents the number of links pointing to a particular Web page or Web site since its creation. The functions offered by search engines, at this moment, do not allow us to retrieve inlinks that were created during a specific period of time. Somehow, this is similar to the nature of financial position variables that accumulates over time. Financial performance measures as revenue or total income also tend to increase over time based on previous performance. According to the evidence found, inlinks seem to reflect mainly the size of a company and to a lower extent, some measures of business performance.

Finally, these studies allow us to say that the significant correlations found in the previous studies are also verified in different industries, out from the Information Technology industry, and in Europe.

4.2.2 Research question 2: Which financial variables explain the inlink count received by a business Web site?

The second research question was answered by designing and testing an explanatory model of the number of inlinks based on financial variables of companies in Spain and the United Kingdom (see section 3.1.3). This is an attempt to explain the inlink variable in commercial Web sites using

a multiple regression analysis. Vaughan and Thelwall (2005) applied a multiple regression analysis to explain the linking pattern in Canadian Universities. The model was tested in five different industries and some general conclusions were found to answer this research question:

The variable “Total assets” (transformed using natural logarithm) is the most relevant variable to explain the number of inlinks received by the business Web sites. Only in the publishing industry this variable is not significant. We consider that this variable should consistently appear in a explanatory model of the number of inlinks received by business Web sites in most industries.

The variable “Intangible assets” is not significant in any industry. There could be several explanations for this: inconsistencies in the way that companies report this information, difficulties by accounting standards to recognize and measure intangible assets and lack of relationship between the intangible assets recognized in the books and the intangible assets of the company related to the Web.

“Turnover” is included as a significant variable for the Construction and the Publishing industries; whereas “Profit after tax” only in the Telecommunications industry. Also the country origin of the company (dummy variable) is significant in 3 out of 5 industries.

To conclude, “Total assets” (variable of financial position) and “Turnover” (variable of performance position) are the most relevant variables that explain the number of inlinks received by commercial Web sites. Also, country origin seems to be a variable to be taken into account, although it is not clear in this study if this variable represents a geographic or a linguistic component, or a mix of both.

4.2.3 Research question 3: How are companies belonging to different industries related when observed through hyperlink structure in the Web?

The third research question was answered by studying several stock exchange indexes including companies belonging to different industries (in section 3.2.1). This research is based on co-link analysis. Web co-link analysis for business information started with a single industry (Vaughan and You, 2006) or on a detailed picture of the competitive landscape of a sector within an industry (Vaughan and You, 2008; Vaughan and You, 2009). The methodology developed in these papers has also been tested and verified in different countries and industries, e.g. China’s chemical industry and electronics industry (Vaughan, Tang and Du, 2009). Parallel to these studies on commercial Websites, research on heterogeneous Web sites has been carried out to test the triple helix theory on the Web (Stuart and Thelwall, 2006; García-Santiago and de Moya-Anegón, 2009). This theory analyzes the transference of knowledge between business, university and government.

This study extended co-link analysis to Web sites of heterogeneous companies belonging to five stock exchange indexes. Multidimensional scaling was used to map the relative positions of the companies in the indexes. We compared results from the five different stock exchange indexes that represent different economic, geographical and cultural backgrounds. There is one main variable that could determine a generalized pattern between economic activities, this is, the degree in which the activity is information centered. Industries whose business model is more information centered form distinct clusters located at a position that is distant from other companies located in more central position. Information centered industries, according to our interpretation, comprise mainly four groups: companies based in the production of information contents (so called Media in the study), companies providing information infrastructures and services (IT), companies very intensive in applying e-commerce techniques (Leisure companies in the tourism industry) and Financial companies. These industries are in an ongoing process of business model transformation, e.g. the deep crisis in the media sector due to digitalization of information.

In addition to their location in the maps, Media and IT companies receive more inlinks than any other companies that belong to more traditional industries. This particular inlink pattern reflects a business model that is highly exposed to Internet. Also, individual maps show that the expectation of finding same industry companies grouped together is only partially fulfilled.

Eurostoxx 50 is the only index in the study that is made of companies belonging to different countries. This fact allowed us to observe economic integration in the Eurozone. Results show that the cluster of companies in this map is better explained by country origin of the company rather than by industries. As a contrast, Dow Jones map is more clearly defined in terms of industries, underlining a deeper economic integration of the U.S. market.

This study reveals that Web co-link data constitute a readily available source of information to obtain new insights into the business environment. The study of heterogeneous companies belonging to different industries allows managers, financial analysts and other stakeholders to locate their company's activity in relation to others and to identify differing business models and to follow their evolution over time. If a company is trying to increase its presence online, inlink count constitutes a more direct measure of the success of the company’s effort in achieving this goal. In addition, co-link analysis could be used by managers to monitor the progresses of the company in relation to other companies in the same industry or in different industries by observing how the co-link MDS maps change. Therefore, co-link analysis could become a managerial tool useful for companies changing their IT strategies in the Web. Conclusions are mainly exploratory but relevant to answer this research question and to open a new direction in the Webometric research of commercial and economic issues.

4.2.4 Research question 4: Can co-link analysis be used to investigate particular economic events?

The forth research question was answered by performing a co-link analysis to study the evolution of the 2009 financial crisis in the U.S. banks listed in the New York Stock Exchange. This research was inspired by a recent study (Vaughan & You, 2008) that proposed a method that combines page content with co-link data to achieve a more detailed picture of the competitive landscape of a sector within an industry. We apply this content assisted method to the banking industry in order to study the impact of the current financial and economic crisis. The keywords selected to refine the search were “crisis”, “bailout” and “subprime”, which are used frequently to describe the current worldwide financial problems. These keywords were intended to filter out pages that link the banks, whether they are inlinks or co-inlinks, for reasons other than the financial crisis. With this study (section 3.2.2), we tried to find out if the inlink and keyword data could provide information on banking financial crisis. Therefore we expected that the data could be used to visualize clusters of banks with more distress. Preliminary results seemed to offer promising results in January 2009. However, a second round of data in August 2009 did not confirm these findings. Some of the reasons that can explain this situation could be the following:

- Co-link analysis could not be an appropriate method to achieve the intended goals. Not all the Web pages linking to a bank with financial distress link simultaneously to another bank in trouble. It seems that a direct link analysis could be a better approach.
- The variable used to measure the financial crisis (the money received by the federal government) is affected by a size effect because largest banks are the banks that logically received more money and therefore it is possible that we are measuring the size of the entity instead of the degree of financial crisis.

Regarding this research question, further research needs to be done in other issues to evaluate the usefulness of this approach. This study can provide guidelines to avoid future problems.

4.2.5 Research question 5: How could different Webometric methods be combined to provide an overall approach to particular industries?

The fifth and final research question was answered by combining the link impact analysis and the co-link analysis within the same paper to analyze a single industry. This study (in section 3.3.1) analyzed the international banking industry through various Webometric techniques. First, we wanted to find out if there is any correlation between the number of Web hyperlinks that a bank attracts and the bank’s financial variables. Second, we used co-link analysis to map these banks’ global market positions. The study achieved its goal of using a combination of inlink and co-link analysis by providing an overview of the global banking industry. We found significant correlations

between inlink counts and various financial variables. A comparison of the inlink count between Asian banks and other banks showed that Asian banks received more inlinks. This is consistent with the fact that Asian banks weathered the recent world financial crisis relatively better.

The multidimensional scaling maps based on co-links indicate that the global banking industry is partially organized in regional markets despite the existence of main global players due to the internationalization process of financial activities. Many of them seem to be more isolated, such as the Chinese banks. This is possibly one of the reasons why the Chinese banks did not suffer from the recent world financial crisis as much as other banks. However, data collected from the second time period revealed changing positions of the Chinese banks. They moved closer to major players such as U.S. and U.K. banks. This reflected the changing perception of the global financial industry toward the Chinese banks as the result of their relative resilience in the world financial crisis.

Although more sophisticated and integrated approaches need to be developed to study companies from different Webometric perspectives, this study shows how different methods could provide complementary views on a particular economic issue.

4.3 Limitations

Although the research provided additional evidence on phenomena previously reported and suggested new lines for future research in business studies using Webometric techniques, there are some limitations in the present research.

Its dynamic nature makes the Web a challenging source of data and this has to be taken into account when interpreting any link based research results. Web research is always a snapshot of that particular time and situation on the Web and this evolves every second. This also makes it impossible to do a research based on linking patterns on the past and to reproduce the same results. Some of the limitations were underlined when we addressed the search engines limitations as a source of data for Webometric research. Therefore, it is also unclear to what extent the results could be generalized when discussing the Web.

The usage of search engines to collect hyperlink information is also affected by the characteristics of the search engine market, which is an oligopoly of three operators Google, Yahoo! and Bing (Microsoft). From a global perspective, they share the majority of the search engine market, although in specific areas of the Web there are other players, for instance, Technorati for searching blogs. Search engine industry is under a constant process of change and innovation. This issue has been treated by many papers in recent years (Lewandowski, Wahlig and Meyer-Bautor, 2006; Evans, 2007). In 2009 Yahoo! and Microsoft sign an agreement that could affect the inlink search functions that so far Yahoo! is providing. This could imply major changes in the development of Webometric research based on search engines data.

Regarding co-link analysis, the studies could present some limitations derived from the interpretation of these results, which could differ depending on the criteria used on the analysis or in the knowledge of the researcher about a specific issue in the study.

4.4 Future research

Although this research answered some questions about commercial Web sites and their inlink patterns, it also creates new questions. The discovery of new evidence on the relationships between inlink counts to business Web sites and financial variables reinforces the evidence previously found (Vaughan, 2004a, 2004b; Vaughan and Wu, 2004). It is necessary to develop models to display this information for business purposes. For example, the elaboration of a model in line with the multiple regression analysis in section 3.1.3 could be used to predict the number of inlinks a Web site is expected to receive based on financial variables and industrial affiliation. Then we could determine which companies are under-performing or over-performing in the Web. Inlink counts, as a measure of the business presence in the Web, could also be used to quantify intangible assets for the company related to the Web.

The use of co-link analysis to study heterogeneous companies in terms of industry provides some promising results, but more research is needed to test the usefulness of the method in order to identify industries that are evolving to business models more intensive in information. In addition, the use of new visualization techniques need to be explored to extract more information from the data.

The use of link impact analysis and co-link analysis have shown promising results for the analysis of business Web sites, however a more systematic approach need to be develop to interpret the results and to combine the methodologies in a way that an overview of the industry is provided. New Webometric measures need also to be explored, such as outlinks or co-outlinks, in the study of commercial Web sites.

Another promising approach is to carry out longitudinal studies to analyze the evolution of industries, stock exchange indexes or single companies at one time. This approach was demonstrated in the studies in section 3.2.2 and 3.3.1.

Qualitative research is a necessary complement to quantitative research because it provides confirming evidence about the relevant nature of the links being analyzed (Vaughan, Gao and Kipp, 2006). This can also reveal the nature of the Web pages, especially to know if they belong to Web 2.0 tools or not. The strong development of the Web 2.0 (blogs, social networking sites, wikis, etc.) in the last five years has changed significantly the Web scenario and its impact can be measured by using a Webometric approach. One of the advantages of studying the Web 2.0 is the

possibility of using alternative sources to collect web data, for instance, *delicious* (a social marking service), *Flickr* (a photography based community), *Youtube* (a video based community), *Technoraty* (a search engine specialized in blogs), etc. The development of the Semantic Web and the use of APIs can widen the research options.

Finally, another suggestion for future research is to apply link impact analysis and co-link analysis to investigate the competitive and cooperative relationships in other areas, such as the public sector or political parties.

4.5 References

Bar-Ilan, J. (2000). Evaluating the stability of the search tools Hotbot and Snap: a case study. *Online Information Review*, 24 (6): 439-449.

Bar-Ilan, J. (2004). The use of Web search engines in information science research. En B. Cronin (Ed.), *Annual review of information science and technology* (pp. 231–288). Medford, NJ: Information Today.

Batelle, J. (2006). *Buscar. Cómo Google y sus rivales han revolucionado los mercados y transformado nuestra cultura*. Barcelona: Ediciones Urano.

Berners-Lee, T. (1999). *Tejiendo la red*. Madrid: Siglo XXI.

Björneborn, L. (2004). *Small-world link structures across an academic Web space: A library and information science approach*. Doctoral dissertation, Royal School of Library and Information Science, Copenhagen, Denmark. Retrieved July 21, 2008, http://vip.db.dk/lb/phd/phd-thesis.pdf

Björneborn, L., & Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1): 65–82.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14): 1216–1227.

Brin S., & Page L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30: 1-7.

Choi, H., & Varian, H. (2009). Predicting the Present with Google Trends. Retrieved May 10, 2009, http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

Evans, M. P. (2007) Analysing Google rankings through search engine optimization data. *Internet Research*, 17(1): 21-37.

García-Santiago, L., & de Moya-Anegón, F. (2009). Using co-outlinks to mine heterogeneous networks. *Scientometrics ,* 79(3): 681-702.

Garfield, E. (1979). *Citation indexing: Its theory and applications in science, technology and the humanities*. New York: Wiley, Interscience.

Gomes, B., & Smith, B.T. (2003). Detecting query-specific duplicate documents. U.S. Patent 6,615,209, [online], retrieved November 15, 2009, http://www.patents.com/Detecting-query-specific-duplicate-documents/US6615209/en-US/

Kahaner, L. (1996). *Competitive Intelligence – How to Gather, Analyze, and Use Information to Move Your Business to the Top*, 7th ed., New York: Touchstone.

Ke, Y., Deng, L., Ng, W., & Lee, D.L. (2006). Web dynamics and their ramifications for the development of Web search engines. *Computer networks*, 50(10): 1430-1447.

Koster, M. (2009a). A standard for robot exclusion. Retrieved February 26, 2010, http://www.robotstxt.org/orig.html

Koster, M. (2009b). About /robots.txt. Retrieved February 26, 2010, http://www.robotstxt.org/robotstxt.html

Lawrence, S., & Giles, C.L. (1998). Searching the World Wide Web. Science, 280:98-100. Retrieved March 3, 2010: http://clgiles.ist.psu.edu/papers/Science-98.pdf

Lawrence, S., & Giles, C.L. (1999). Searching the Web: General and scientific information access. IEEE Communications, 37(1): 116-122.

Lewandowski, D., Wahlig, H., & Meyer-Bautor, G. (2006) The freshness of web search engine databases. Journal of Information Science, 32 (2): 131–148

Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines—Fluctuations in document accessibility. *Journal of Documentation*, 57(5): 623–651.

Porter, M.E. (1980) *Competitive Strategy: Techniques for Analyzing Industries and Competitors.* New York, NY: Free Press.

Porter, M.E. (1985). *The Competitive Advantage: Creating and Sustaining Superior Performance*. New York, NY: Free Press.

Reid, E. (2003). Using Web link analysis to detect and analyze hidden Web communities. En: Vriens, D. (ed.), *Information and communications technology for competitive* intelligence (pp. 57-84). Hilliard, OH: Ideal Group.

Shaw, D. (2001). Playing the links: interactivity and stickiness in .com and “not.com” websites. *First Monday*, 6(3). Retrieved May 10, 2009, http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/837/746

Sherman, C., & Price, G. (2001). *The invisible Web*. Medford, NJ: Information Today, Inc.

Small, H. (1973). Co-citation in the scientific literature: a new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4): 265-269.

Smith, A., & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics*, 54: 363-380.

Stuart, D., & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: a case study of the UK West Midlands automobile industry. *Research Evaluation*, 5(2): 97–106.

Tan, B., Foo, S., & Hui, S. C. (2002). Web information monitoring for competitive intelligence. *Cybernetics and Systems*, 33(3): 225-251.

Thelwall, M. (2001a). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52(13): 1157–1168.

Thelwall, M. (2001b). The responsiveness of search engine indexes. Cybermetrics, 5(1), paper 1. Retrieved February 26, 2009, http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html

Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology*, 59(1): 38-50.

Thelwall, M. (2009) *Introduction to Webometrics. Quantitative Web Research for the Social Sciences*, Morgan & Claypool.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. En B. Cronin (ed.) *Annual review of information science and technology* (pp. 81–135). Medford, NJ: Information Today.

Turow, J., & Tsui, L., Editors (2008). The Hyperlinked Society: Questioning Connections in the Digital Age. Ann Arbor: University of Michigan Press and University of Michigan Library. Retrieved February 18, 2010, http://quod.lib.umich.edu/cgi/t/text/text-idx?c=nmw;;idno=5680986.0001.001

Uyar, A. (2009). Investigation of the accuracy of search engine hit counts. *Journal of Information Science*, 35(4): 469-480.

Vaughan, L. (2004a) Exploring website features for business information. *Scientometrics*, 61(3): 467–477.

Vaughan, L. (2004b). Web hyperlinks reflect business performance—A study of US and Chinese IT companies. *Canadian Journal of Information and Library Science*, 28(1): 17–31.

Vaughan, L. (2006). Visualizing Linguistic and Cultural Differences Using Web Co-Link Data. *Journal of the American Society for Information Science and Technology*, 57(9): 1178-1193.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal web sites? *Journal of the American Society for Information Science and Technology*, 54:

29-38.

Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*, 40(4): 693-707.

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

Vaughan, L., & Wu, G.Z. (2004). Links to commercial websites as a source of business information. *Scientometrics*, 60(3): 487–496.

Vaughan, L., & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3): 611–628.

Vaughan, L., & You, J. (2008). Content assisted web co-link analysis for competitive intelligence. *Scientometrics*, 77(3): 433-444.

Vaughan, L., & You, J. (2009). Keyword enhanced Web structure mining for business intelligence. *Lecture Notes in Computer Science*, 4879: 161–168.

Vaughan, L., & Zhang, Y. (2007). Equal representation by search engines? A comparison of Websites across countries and domains. *Journal of Computer-Mediated Communication*, 12(3), retrieved February 26, 2010, http://jcmc.indiana.edu:80/vol12/issue3/vaughan.html

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

Vaughan, L., Tang, J., & Du, J. (2009). Examining the robustness of Web co-link analysis. *Online Information Review*, 33, (5): 956-972.

Zuccala, A. (2006). Author Cocitation Analysis Is to Intellectual Structure A Web Colink Analysis is to...? *Journal of the American Society for Information Science and Technology*, 57(11): 1487-1502.