



# Decomposing social and semantic networks in emerging “big data” research



Han Woo Park<sup>a,\*</sup>, Loet Leydesdorff<sup>b</sup>

<sup>a</sup> Department of Media and Communication, Yeungnam University, South Korea

<sup>b</sup> Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Kloveniersburgwal 48, 1012 CX Amsterdam, The Netherlands

## ARTICLE INFO

### Article history:

Received 9 January 2013

Received in revised form 30 May 2013

Accepted 30 May 2013

Available online 20 July 2013

### Keywords:

Data science

Big data

International co-authorship

Social network analysis

SCI

Semantic network

## ABSTRACT

This paper examines the structural patterns of networks of internationally co-authored SCI papers in the domain of research driven by big data and provides an empirical analysis of semantic patterns of paper titles. The results based on data collected from the DVD version of the 2011 SCI database identify the U.S. as the most central country, followed by the U.K., Germany, France, Italy, Australia, the Netherlands, Canada, and Spain, in that order. However, some countries (e.g., Portugal) with low degree centrality occupied relatively central positions in terms of betweenness centrality. The results of the semantic network analysis suggest that internationally co-authored papers tend to focus on primary technologies, particularly in terms of programming and related database issues. The results show that a combination of words and locations can provide a richer representation of an emerging field of science than the sum of the two separate representations.

© 2013 Elsevier Ltd. All rights reserved.

## 1. Introduction

There has been growing interest among scientists, policymakers, industry professionals, and journalists in the development of research driven by big data as technologies and practices for research and development worldwide (Halevi & Moed, 2012; O'Reilly Radar Team, 2012; OECD, 2012; Rousseau, 2012). The term “big data” refers to “analytical technologies that have existed for years but can now be applied faster, on a greater scale and are accessible to more users.” (Miller, 2013, p. 1). In this regard, research driven by big data reflects a discipline that, to extract meaning from very large datasets, incorporates various techniques such as data mining and visualization into diverse fields, including the humanities and social sciences. In a special issue published on February 11, 2011, *Science* looked broadly at research efforts driven increasingly by big data as a scientific domain (Science, 2011). Several prominent institutions have organized international conferences, for example, “IPP 2012: Big Data, Big Challenges” conference by the University of Oxford (<http://blogs.oii.ox.ac.uk/policy/?p=447>).

Because the most noticeable thread running through national competitiveness in an increasingly digital society is to cope with the scalability of data, research driven by big data has become a priority in both academia (Harris, 2012; King, 2011) and industries (McKinsey Global Institute, 2011; Rogers, 2011). Because big data can have substantial impact on various aspects of society, the “Big Data Public Private Forum” was also established in order to offer a major boost for supporting actions for the successful implementation of “the big data economy” (<http://www.big-project.eu/>). According to an EMC report on global big data (<http://www.emc.com/about/news/press/2012/20121211-01.htm>), Western Europe is currently making the

\* Corresponding author at: Department of Media and Communication, YeungNam University, 214-1 Dae-dong, Gyeongsan-si, Gyeongsangbuk-do 712-749, South Korea. Tel.: +82 53 810 2275; fax: +82 53 810 2270.

E-mail addresses: [hanpark@ynu.ac.kr](mailto:hanpark@ynu.ac.kr), [parkhanwoo@hotmail.com](mailto:parkhanwoo@hotmail.com) (H.W. Park), [loet@leydesdorff.net](mailto:loet@leydesdorff.net) (L. Leydesdorff).

heaviest investment in managing the deluge of digital data, spending \$2.49 per gigabyte (GB), followed by the U.S. (\$1.77), China (\$1.31), and India (\$0.87).

This study investigates the visibility of countries in terms of their research collaboration in the domain of big data in the context of scientific documents. The study examines the use of related terms by country based on social and semantic network methods and provides a quantitative analysis of how various terms are communicated in academia. More specifically, the study provides an empirical analysis using SCI (Science Citation Index) journals for the year 2011. We consider the following research questions:

**RQ1.** What is the structural pattern of the network of internationally co-authored SCI papers in the domain of research driven by big data? Which countries occupy central or peripheral positions in the 2011 network?

**RQ2.** What is the semantic structure of paper titles in research driven by big data?

## 2. Literature review

This section provides a theoretical introduction based on a literature review. Big data research is composed of interrelated clusters of tasks. For example, technologies for accessing, collecting, and organizing large amounts of data and unique skill sets have played an increasingly central role in big data research (Loukides, 2012). Rousseau (2012) claimed that the definition of “large amounts of data,” i.e., big data, depends on concrete circumstances related to individual projects. It may mean tens of terabytes ( $10^{12} \approx 2^{40}$  bytes) for some research, whereas it may mean several petabytes ( $10^{15} \approx 2^{50}$  bytes) or even exabytes ( $10^{18} \approx 2^{60}$  bytes) for large enterprises and huge research projects. From this perspective, Loukides (2012) claimed that today’s “big” is probably tomorrow’s “medium” and next week’s “small” and thus that the most effective definition of “big data” may be derived when the size of data itself becomes part of the research problem.

The origin of big data research is rooted mainly in various initiatives started in the early 1970s. One such initiative includes Peter Naur’s 1974 book “Concise Survey of Computer Methods,” a survey of contemporary data-processing methods in a wide range of applications (Gilpress, 2012). During the 1990s, the term was associated mainly with computer modeling and software development for large data sets. The launch of the journal *Knowledge Discovery and Data Mining* in 1997 represents such a trend in computer science. Rousseau (2012) regarded the early 1990s as the beginning of the field of big data research. Recent years have witnessed the emergence of scholarly journals focusing mainly on research driven by big data, including *Data Science Journal* (2002), *Journal of Data Science* (2003), *EPJ Data Science* (2012), and *Journal of Big Data* (2013). While these journals address the general aspects of big data research, *GigaScience*, established in 2012, focuses specifically on the study of “big data” from the entire spectrum of life and biomedical sciences.

Although research based on big data can be conducted in various ways, its basic purpose lies in handling huge amounts of data from technological, sociological, and economic systems to discover some hidden patterns. This is referred to as the “fourth paradigm,” which has been described as a new approach to scientific exploration. This scientific approach is attributed to Jim Gray, one of the most influential computer scientists at Microsoft (Hey, Tansley, & Tolle, 2009). The establishment of big data research as a complementary approach to traditional hypothesis-driven methods started in natural and engineering domains and later moved to social and behavioral sectors in the mid-2000s (Bonacich, 2004). Savage and Burrows (2007) stated that big data research is closely associated with changes in conducting scientific tasks and lamented that “fifty years ago, academic social scientists might be seen as occupying the apex of the—generally limited—social science research ‘apparatus’. Now they occupy an increasingly marginal position in the huge research infrastructure.” (p. 886).

In particular, the rapid development of mapping-oriented visualization and e-research practices, together with the universality of advanced research networks and cloud computing, has facilitated the dissemination of big data research in the late 2000s (Jankowski, 2009). The volume of data available for research has suddenly become controversial since the “end of theory” claim by Anderson (2008), who argued that data are everything researchers need and thus that they do not have to settle for models. Knowledge can be discovered through correlations (instead of causality) between big data and related events/observations, experiments, and simulations by enriching traditional data in novel ways, posing a serious challenge to social scientists. In response, Graham (2012) suggested that “We may one day get to the point where sufficient quantities of big data can be harvested to answer all of the social questions that most concern us. I doubt it though. There will always be digital divides; always be uneven data shadows; and always be biases in how information and technology are used and produced.” Both the definition and implementation of big data research may still need to be clarified for its successful adoption in a society. However, several researchers have become increasingly concerned with big gaps of knowledge in the changing research ecology made possible by big data (Snijders, Matzat, & Reips, 2012). The creation, organization, and analysis of big data can be shaped by a number of interdependent social, cultural, and political subsystems in the research ecosystem. To reduce the barrier to entry into big data research, the OECD initially suggested an international framework in the following five domains for open access to scholarly dataware houses ranging from gigabytes to petabytes of data (Arzberger et al., 2004a, 2004b): technological, institutional/managerial, financial/budgetary, legal/policy, and cultural/behavioral domains. The OECD (2012) recently organized a technology foresight forum in Paris (“Big Data Analytics and Policies”) to harness big data as a new source of growth. In the Europe 2020 strategy, the European Commission identified the use of big data as a new driver of growth and jobs under seven flagship initiatives (<http://ec.europa.eu/europe2020>).

Because data streams from many sources can be overwhelming, the ubiquitous use of big data across various segments of society does not necessarily mean intellectual coherence. Halevi and Moed (2012) examined the development of research related to big data by using the Scopus database. They searched for the term “big data” by using the index and author keyword fields from 1970 to early 2012. An examination of 306 documents revealed a sharp increase in such research in conference proceedings (131 papers), particularly in computer science (171 documents). In a similar vein, Rousseau (2012) collected 141 documents using the web version of the Web of Science database based on TS (topic search) = “big data” and found the salience of conference proceedings (54 papers vs. 47 articles) in computer-related fields from 1993 to October 2012. Both studies concluded that this research was led mainly by computer science in the early years (until 2006) but spread to other engineering and natural science fields in recent years. Further, the U.S., China, and Germany were the most active countries in terms of the publication of papers based on big data.

Previous studies have not systematically examined whether research efforts driven by various sources of big data are really becoming increasingly widespread across the world. Further, the status of the literature based on big data has not been extensively discussed or sufficiently examined with respect to its semantic variations, disciplinary scope, institutional adoption, and international collaboration. For example, Halevi and Moed (2012) and Rousseau (2012) focused on descriptive statistics. Therefore, the present study incorporates a network perspective into both social (in terms of co-authorship) and semantic networks. In addition, the study extends search queries to various terminologies related to research enabling big data because the term “big data” is regarded as only one of many terms. The study shows where the research system focusing on big data is particularly popular in terms of international collaboration and prevailing semantics.

### 3. Methods: data collection and analysis procedures

#### 3.1. Social network analysis and the co-authorship network

We employed a method rooted in social network analysis (SNA) (for a detailed explanation of social network metrics and technical procedures, see Hanneman & Riddle, 2005). Here the unit of analysis is often the node, which refers to a point in a network where ties cross or connect nodes. A tie is a connection between parts (i.e., nodes) in a network. In the present study, we considered countries as nodes and a tie as the number of papers co-authored by a pair of researchers with different addresses in terms of their country of origin. Previous studies (Kwon, Park, So, & Leydesdorff, 2012; Otte & Rousseau, 2002; Park & Leydesdorff, 2008, 2010; Shapiro & Park, 2012) have shown that indicators and visualizations based on SNA are particularly useful for identifying the structural pattern of research collaboration between a set of authors. In addition, a “new invisible college” embedded in scientific networks can be disclosed (Wagner, 2008).

#### 3.2. Data collection

For the data, we considered papers published in SCI journals in 2011. For this, we selected three types of documents: journal articles, letters, and reviews. We obtained the data from the DVD version of the SCI database by using several search terms based on titles, author keywords, and keyword-plus. In addition, we considered the space between two words. For example, we used the terms “big data” and “bigdata” as a search string.

We made search strings based on the following procedure: First, we attempted to find some authoritative sources listing all relevant keywords, but to the authors' knowledge, no related studies provide a solid list of queries. Therefore, we required a more objective technique. Second, we used the term “big data” as a seed search string for “titles,” “author keywords,” and “keyword plus” to retrieve relevant documents from the DVD version of the SCI 2011 database. However, we retrieved only four documents with the term “big data.” We considered including some terms co-occurring with “big data,” including “cyberinfrastructure,” “data mining,” and “clouds.” To complement the SCI DVD, we also considered the SCI(E) 2011 database from the Web of Science (WoS) database by using Topic = (“Big Data”) as the search string and retrieved a total of 15 documents. The important terms related to “big data” included “data integration” and “computational science.” Third, the term “big data” in the title provided 889 documents through Google Scholar. We included the most frequently co-occurring title words with “big data” in this set, including “clouds,” “cloud computing,” “Hadoop,” “MapReduce,” “big science,” “NoSQL,” and “data warehouse.” Therefore, we used a total of 11 keywords (“big data,” “big science,” “cloud computing,” “computational science,” “cyberinfrastructure,” “data integration,” “data mining,” “data warehouse,” “Hadoop,” “MapReduce,” and “NoSQL”) for the final search string.

Before the finalization of search queries, we personally sent a list of potential queries to Rousseau (personal communication, February 27, 2013), who provided a scientometric mapping of the development of the domain of big data research by using the web version of the bibliometric database (Rousseau, 2012). During the consultation, we realized that the term “parallel computing” has little to do with big data research because this technique has been studied and applied in the last few decades. Finally, we removed duplicate documents collected from a total of 477 records by using 11 search strings and thus had a total of 406 records for the final analysis.

#### 3.3. Analysis methods and indicators

Using a list of scientific documents selected as relevant to big data science, we created a binary matrix containing 406 records of documents (articles, reviews, and letters) as rows and countries as variables (i.e., columns). Here

if one country has four addresses listed and another country, two, then some social network analysis tools count them as  $4 \times 2 = 8$  collaborations. In the case of binary values, there is a single ( $1 \times 1$ ) collaboration between these two countries. Given the research questions, we used this binary file. We considered as missing records those with no addresses.

We used this asymmetric binary matrix in UCINET to generate a symmetrical and valued co-authorship network. We first examined the structural properties of this co-authorship network in the field of big data science relative to those of the global co-authorship network in 2011 (Leydesdorff, Wagner, Park, & Adams, 2013). After the normalization, we used the global distribution to specify expected numbers for a more in-depth analysis.

Structural properties can be detected in several ways. A major indicator is network density, which is the total number of ties divided by the total number of possible ties. For a valued network, this index is the total of all values divided by the number of possible ties (Borgatti, Everett, & Freeman, 2002). Along with density, *centralization* is a useful measure for describing how densely the ties are concentrated around a focal node in a given network. This is expressed as a percentage with respect to the most central or unequal network possible for any number of nodes (Hanneman & Riddle, 2005). For example, in a “star” network, there is only one node connected to all other nodes in the network. Therefore, network centrality is 100% because in such a network, all nodes are connected only to the central node.

To calculate the correlations between the global network and its subset limited to big data science, we employed the quadratic assignment procedure (QAP) in UCINET. The Pearson correlation in the QAP is regarded as a standard measure when both networks have valued (non-binary) relationships measured at the interval level (Hanneman & Riddle, 2005). Here a low probability ( $p < 0.05$ ) suggests a fairly strong relationship between matrices that is unlikely to have occurred by chance.

In addition, we statistically tested the international co-authorship matrix for any significant differences from the expectations (based on the whole database). As suggested by Leydesdorff and Bornmann (2012), the statistical significance of performance differences between two proportions can be tested using the z-test for independent proportions. An absolute value of z greater than 1.96 indicates a significance difference between two networks at the 5% level ( $p < 0.05$ ). Here the critical values for a test at the 1% level ( $p < 0.01$ ) and the 0.1% level are  $|z| > 2.576$  and greater than 3.29, respectively (<http://easycalculation.com/statistics/p-value-for-z-score.php>). For a more rigorous conclusion, we used the most demanding significance level ( $p < 0.001$ ).

We considered degree centrality to measure the number of countries with which a country engaged in direct co-authorship collaboration (Park & Leydesdorff, 2008). Here the more a node (in this study, a country) is connected directly to other nodes in the network, the more central the node is (for a detailed calculation of network metrics, see Wasserman & Faust, 1994). This means that researchers in a central country are more likely to have opportunities for co-authorship collaboration (Hanneman & Riddle, 2005). We used degree centrality to measure the direct relationship between a country and its neighbors in a co-authorship network (e.g., reputation, popularity, activity, and prominence in the flow of information).

In addition to degree centrality, betweenness centrality measures the strategic position of the country in terms of coordinating scientific collaboration between countries (Leydesdorff, 2007; Park, 2012). There are two major indicators of betweenness centrality: node and flow betweenness. The node betweenness centrality of a country refers to the proportion of all geodesic relationships between pairs of other nodes (in this study, countries) that pass through a given node (Freeman, 1979). In other words, betweenness centrality refers to the extent to which a node lies along the shortest path connecting others in the network. Freeman, Borgatti, and White (1991) expanded betweenness centrality by considering the number of ties (i.e., the strength of ties connecting two nodes in a network) between pairs of other nodes that pass through a node for some maximum flow. For a node, flow betweenness centrality sums up how involved that node is in all the flow between all other node pairs (Hanneman & Riddle, 2005). Because an international co-authorship network is a valued symmetric graph only with integer values, flow betweenness centrality can thus be a good measure of a node's contribution to all possible maximum flows (Borgatti, Everett & Freeman, 2002).

Finally, we calculated “structural holes” (Burt, 1992) to estimate the strategic management of a country in co-authorship networks. From Burt's perspective, one must form a tie with another that has strong relationships with others with whom one does not yet have ties. Selecting such a collaborator reduces communication redundancy and increases networking efficiency in terms of one's own ego network. In the analysis of co-authorship networks in information and library science fields, Abbasi, Chung, and Hossain (2012) associated a researcher's structural hole with their scholarly performance and found that researchers who engage in strong collaboration with one primary co-author within a group of researchers for joint publication perform better than those with many overlapping relationships within the same group of co-authors. We calculated three measures of Burt's structural holes (Burt, 1992). “Network effectiveness” measures the effective size of the ego network. In this case, effective size denotes the number of non-redundant countries in the international co-authorship network. “Network efficiency” is derived by dividing effective size by the number of countries in the ego network. “Network constraints” measure the extent to which one's opportunities (e.g. time and resources that can be used for forming new ties and seeking information) are limited by being connected to others who are already connected to one another. We employed UCINET mainly for social network measures (e.g., density, centrality, and sub-clusters) for the network (Borgatti, Everett & Freeman, 2002).

**Table 1**

A comparison of density and centralization values.

Network type	Density (S.D.)	Centralization (%)		
		Degree	Node	Flow
Global	26.71 (245.70)	5.11	10.08	9.83
Big data	0.01 (0.18)	4.37	2.70	2.28

N = 201.

## 4. Results

### 4.1. RQ 1 – international co-authorship network

We included a total of 201 countries in the 2011 co-authorship network for the analysis of big data science. We aggregated addresses in England, Scotland, Wales, and Northern Ireland as U.K. addresses. Leydesdorff, Wagner, Park, and Adams (2013) provided a detailed discussion on the whole structure of the 2011 co-authorship network at the global level. This section discusses the changes in the network structure across different matrices (Table 1). As expected, the global co-authorship network was far denser than the subnetwork, that is, co-authorship in big data research. Note that these were not really co-authorship relationships between countries but relationships between them measured in terms of co-authorship relationships. The sum of ties in the global network and that of the subnetwork were 1,073,764 and 10,798, respectively. In addition, the global network was more centralized around hub countries than the network of big data science in terms of all three measures of centrality. However, the QAP correlation between the whole 2011 co-authorship network and big data research demonstrates their significant relationship: this (Pearson) correlation was .740 ( $p < .001$ ).

To determine the difference between the big data science network and the grand matrix of all co-authorship relationships, we conducted a z-test. The results show that some countries were overrepresented in the co-authorship network in data science, whereas others, underrepresented. A total of 30 countries had a test value greater than |1.96| ( $p < 0.05$ ). Among these, 22 had positive z-scores, indicating that group 1 (in this case, the co-authorship network in big data science) showed higher frequency than group 2 (in this case, the whole co-authorship network). Among these 22 countries, 50.0% (11) showed z-scores greater than |3.29| ( $p < 0.001$ ), and 27.3% (6), greater than |2.576| ( $p < 0.01$ ). Macedonia showed the highest z-scores, but this may be due to a low expected value. Contrary to expectations, two of Macedonia's 205 international collaborations were in the set of big data science. Several European countries (Denmark, Ireland, Portugal, Norway, and the Netherlands) engaged in international collaboration more than they were expected to ( $p < 0.05$ ). This was likely due to an E.U. policy, but the results provide no clear evidence.

There were some developing and transitional countries, including Taiwan, Singapore, and South Korea, in the group that scored higher than expected. On the other hand, a total of six countries had negative z-scores. None of these countries had z-scores less than 3.29 ( $p < 0.001$ ); three, less than 2.576 ( $p < 0.01$ ); and five, less than 1.96 ( $p < 0.05$ ). Another noteworthy result is that Russia, India, and Japan were underrepresented in international collaboration in big data science at the 5% level. Wagner, Leydesdorff, Adams, and Park (2013) provided a detailed discussion on the 2011 global co-authorship network in terms of science and technology policies.

**Table 2**

Centrality values for countries.

Rank	Country	Degree	Rank	Country	Betweenness	Rank	Country	FlowBet
1	U.S.	4.450	1	U.S.	2.734	1	USA	2.309
2	GERMANY	1.650	2	FRANCE	1.253	2	FRANCE	0.929
3	U.K.	1.600	3	U.K.	0.680	3	CANADA	0.537
4	FRANCE	1.400	4	CANADA	0.643	4	ITALY	0.510
5	AUSTRALIA	1.150	5	ITALY	0.620	5	UK	0.377
6	NETHERLAND	1.150	6	AUSTRALIA	0.602	6	SOUTH_KORE	0.359
7	CHINA	1.100	7	SOUTH_KORE	0.346	7	BELGIUM	0.331
8	DENMARK	0.950	8	GERMANY	0.291	8	AUSTRALIA	0.328
9	CANADA	0.900	9	BELGIUM	0.290	9	JAPAN	0.262
10	TAIWAN	0.850	10	PORTUGAL	0.266	10	SLOVENIA	0.200
11	ISRAEL	0.750	11	JAPAN	0.256	11	PORTUGAL	0.185
12	SOUTH_KORE	0.750	12	CHINA	0.137	12	CHINA	0.132
13	SWEDEN	0.750	13	NETHERLAND	0.104	13	SPAIN	0.129
14	ITALY	0.700	14	DENMARK	0.099	14	GERMANY	0.108
15	PORTUGAL	0.700	15	SAUDI_ARAB	0.088	15	MALAYSIA	0.103
16	IRELAND	0.650	16	SLOVENIA	0.068	16	TANZANIA	0.095
17	NORWAY	0.650	17	TAIWAN	0.057	17	VENEZUELA	0.095
18	SPAIN	0.650	18	SPAIN	0.055	18	NETHERLAND	0.089
19	SINGAPORE	0.500	19	ISRAEL	0.037	19	SAUDI_ARAB	0.071
20	SWITZERLAN	0.450	20	AUSTRIA	0.036	20	AUSTRIA	0.063



**Table 3**  
Structural hole values by country.

Rank	Country	Effectiveness	Rank	Country	Efficiency	Rank	Country	Constrain
1	U.K.	13.071	1	EGYPT	1.000	1	DENMARK	0.312
2	AUSTRALIA	12.879	2	INDIA	1.000	2	NETHERLAND	0.331
3	FRANCE	12.562	3	POLAND	1.000	3	PORTUGAL	0.338
4	U.S.	11.563	4	UZBEKISTAN	1.000	4	ISRAEL	0.343
5	GERMANY	10.746	5	GREECE	0.805	5	NORWAY	0.345
6	NETHERLAND	8.873	6	JAPAN	0.789	6	IRELAND	0.352
7	DENMARK	8.530	7	AUSTRIA	0.725	7	UK	0.364
8	PORTUGAL	8.229	8	BRAZIL	0.722	8	SWEDEN	0.365
9	ISRAEL	8.208	9	NEW_ZEALAN	0.722	9	AUSTRALIA	0.381
10	CANADA	7.672	10	MALAYSIA	0.698	10	GERMANY	0.397
11	ITALY	7.554	11	AUSTRALIA	0.678	11	FRANCE	0.411
12	IRELAND	7.252	12	SAUDI_ARAB	0.667	12	CANADA	0.532
13	NORWAY	7.214	13	IRAN	0.667	13	ITALY	0.535
14	SOUTH_KORE	6.365	14	THAILAND	0.667	14	SAUDI_ARAB	0.548
15	CHINA	6.057	15	SINGAPORE	0.659	15	SWITZERLAN	0.556
16	SWEDEN	5.978	16	CZECH_REPU	0.644	16	USA	0.573
17	JAPAN	5.520	17	CANADA	0.639	17	SOUTH_KORE	0.578
18	TAIWAN	5.490	18	SLOVENIA	0.638	18	BELGIUM	0.583
19	SPAIN	5.312	19	SOUTH_KORE	0.636	19	SPAIN	0.625
20	SWITZERLAN	4.224	20	PORTUGAL	0.633	20	TAIWAN	0.627

4.1.1. Degree and betweenness centrality

The U.S. was the most central country (Table 2) in terms of degree centrality, acting as the network’s hub in research involving big data. The U.S. had the largest number of papers co-authored with other countries, and its degree centrality was 4.450, accounting for 17.5% of the network, followed by Germany, the U.K., France, Australia, the Netherlands, China, Denmark, and Canada, in that order. In addition to China, some other countries in Asia were part of the top 20 collaborators, including Taiwan and South Korea. Table 2 lists the top 20 countries in terms of their nodal and flow betweenness. Canada, Italy, South Korea, and Japan performed better than their degree centrality. Therefore, these countries, with their high betweenness centrality, exerted control over scientific communication as broker or gatekeepers.

4.1.2. Structural holes

Table 3 summarizes some measures of the “structural hole.” A closer examination of structural holes reveals noteworthy results for the top 20 countries across the three measures. Central countries in terms of their degree and betweenness

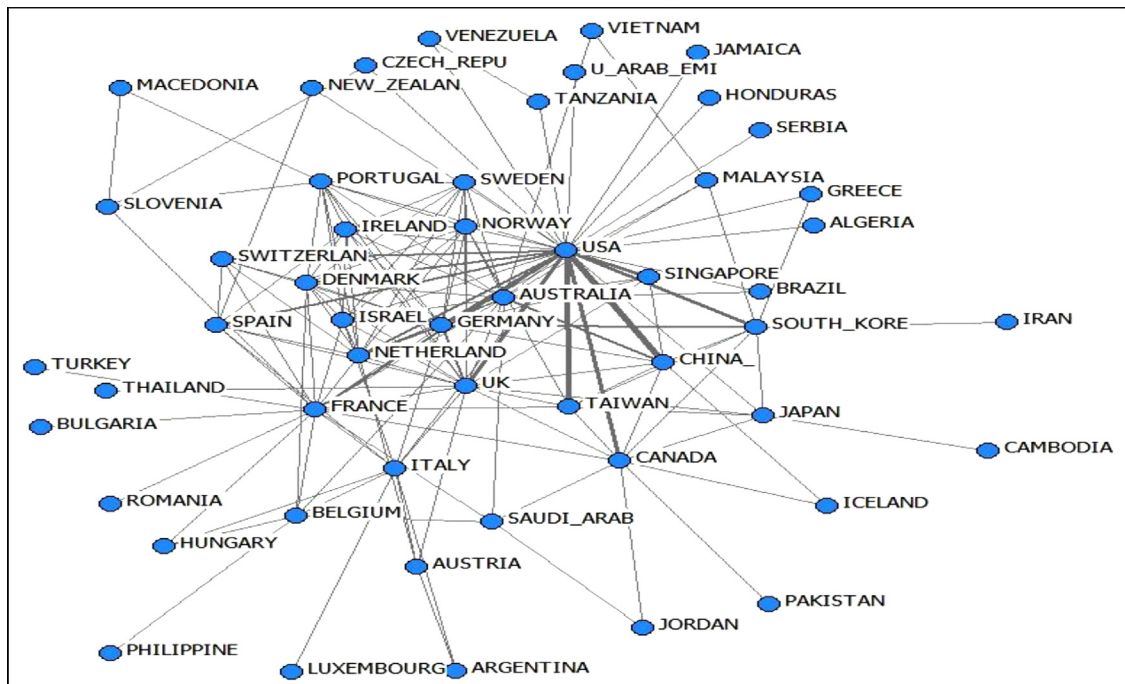


Fig. 1. International co-authorship network of big data research.

centrality also occupied central positions in terms of network effectiveness. However, a majority of these countries are not listed in terms of their network efficiency. In addition, European countries rank high in terms of the constraint indicator. Unexpectedly, the U.S. showed very low efficiency (0.361, 51st) and moderately low constraint (0.573, 16th) indices. This suggests that U.S. researchers may not be strategically positioned for generating new ideas and enjoying greater career mobility at the international level and perhaps that the domestic market is more important for these scholars.

As shown in Fig. 1, there were several countries connected to the U.S. that had the highest degree and betweenness centrality values. Although the U.S. had some strong ties with Asian countries (e.g., Singapore, South Korea, China, and Taiwan), it had a relatively sparse network with European countries (e.g., Switzerland, Denmark, Spain, the Netherlands, and France). Note that the thickness of a line indicates the frequency of co-authorship relationships between two countries. The visualization is based on Kamada and Kawai's (1989) spring-embedding algorithm, in which the closer a pair of countries is in the two-dimensional space, the stronger their relationship. The line length does not directly represent the total number of co-authored papers, but because of the Kamada–Kawai algorithm, the shorter the line, the higher the number of these papers.

## 4.2. RQ2 – semantic network analysis

### 4.2.1. Semantic structure of paper titles

Before visualizing the semantic network, we calculated the frequency of specific terms included in paper titles (for a detailed explanation of the semantic network procedure, see Cho, Choi, & Park, 2012). The 10 most frequently occurring terms included “data mining” (76 times), “data” (56), “using” (54), “analysis” (29), “mining” (28), “application” (27), “cloud computing” (24), “based” (23), “algorithm” (20), and “approach” (20). In addition, 11 terms occurred fewer than 20 times: “cyberinfrastructure” (13), “data integration” (12), “MapReduce” (7), “data warehouse” (5), “computational science” (4), “big data” (2), and “big science” (1). No title included the terms “Hadoop” and “NoSql.”

Fig. 2 shows a semantic network using the top 50 terms and four terms (“data warehouse,” “computational science,” “big data,” and “big science”) used for retrieving papers related to big data. We organized the layout by using the

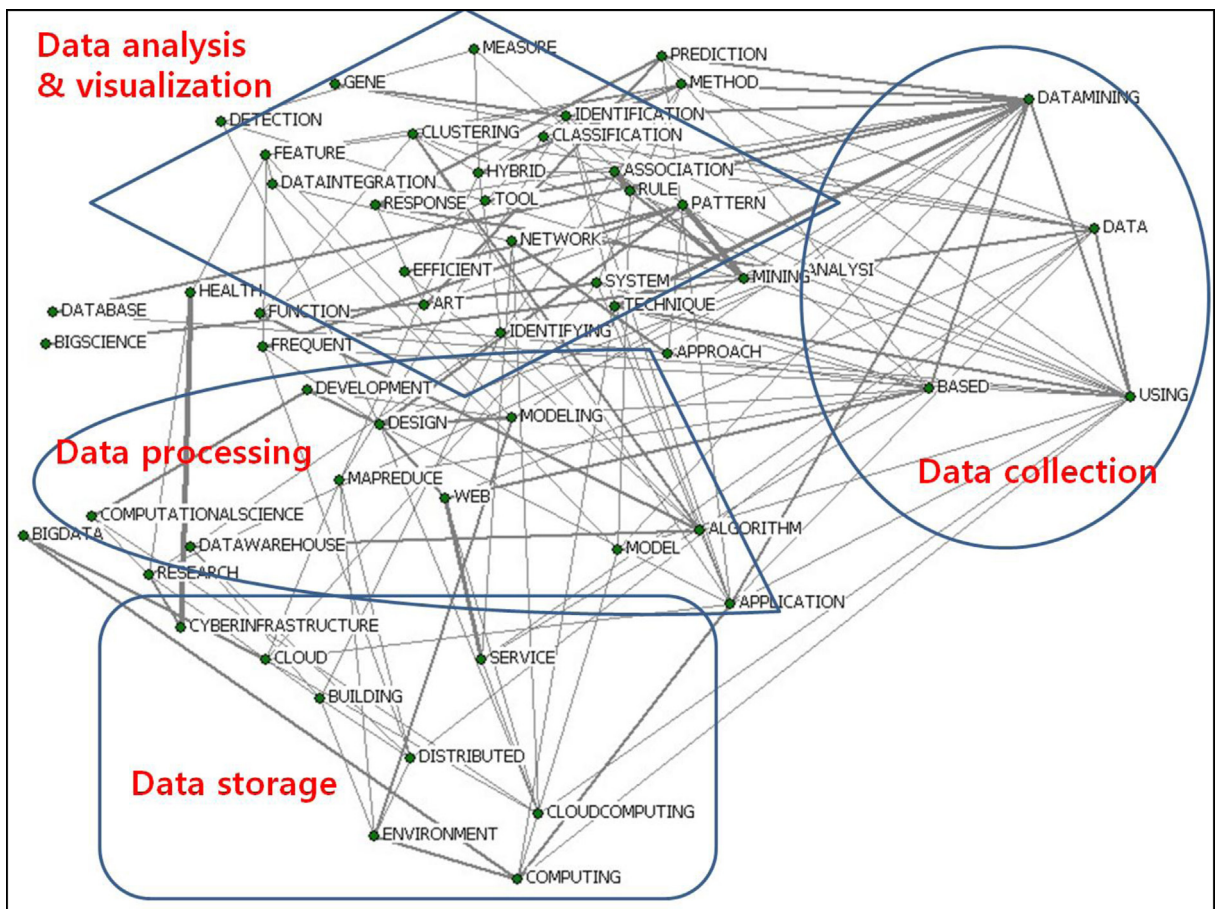


Fig. 2. Semantic network of paper titles in big data (50 most frequently occurring terms with the cosine  $\geq 0.1$ ).

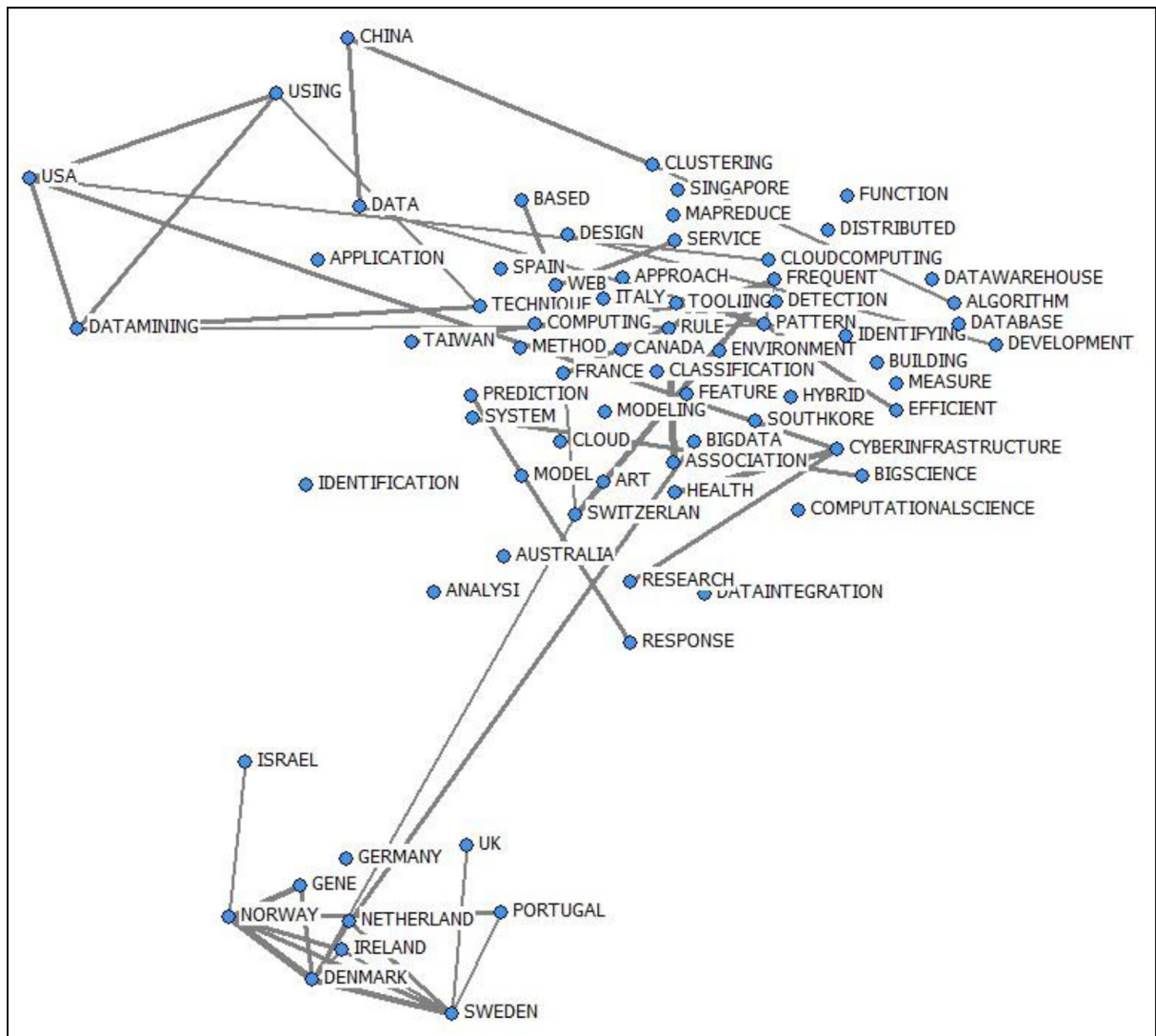


Fig. 3. Semantic network of paper titles and countries in big data (50 most frequently occurring terms and the top 20 countries with the cosine  $\geq 0.2$ ).

principal-component option in NetDraw (Borgatti, 2012). The thickness of a line indicates the strength of semantic connections (in terms of co-occurrences) between two terms in a document. Here we normalized this thickness by using the cosine of concurrence frequency.

As shown in Fig. 3, we used both the 50 most frequently occurring terms and the top-20 countries in terms of degree centrality for the visualization of the semantic and country network in a cosine-normalized network (Leydesdorff, 2010a, 2010b). In the top left corner, the U.S. stands alone by being connected to the most frequently used term – “data mining” – and China is also visible by being connected to the second most frequently occurring term – “data.” Noteworthy is that countries in Northern Europe, including Norway, Denmark, and Sweden, formed a group with the U.K. and other European countries. The prominence of the very specific term “gene” in this cluster reflects a general preference among European researchers (e.g., bioinformatics) and the direction of national policies. In the right corner, there is a large cluster in which several terms and countries are weakly connected.

## 5. Discussion and conclusions

The present study provides a snapshot of any structural changes in the academic community during the transition from theory-based knowledge discovery to research practices driven by big data. Several studies have indicated some changes in the community with respect to research based on big data during the transition period from traditional computer science fields to other engineering, social, and natural science fields. However, internationally co-authored papers in the field of big data research have generally focused on primary technologies (e.g., “data mining”) applying classical techniques to data sets



or to programming and related database issues. This suggests that SCI papers do not necessarily focus on conceptually new methodologies for analyzing and synthesizing massive data sets. However, the results suggest the emergence of some new subjects such as “MapReduce,” implying that researchers are gradually focusing on methodologies that make use of new digital and/or statistical tools to manage the deluge of data.

Although we invested in sharpening the search, one can still doubt whether we were able to retrieve this new field using our search terms. Do the results meaningfully describe an area of science? Emerging fields have both geographic and cognitive components. As shown in the geographic map (Fig. 1), the U.S. was central in various aspects because of its connections with E.U. member countries as well as individual Asian countries. The network shows various European countries with the second most central positions based on centrality measures. In terms of structural hole indicators, however, some smaller and less advanced countries were more “efficient” than “effective” in terms of controlling central positions. This may be due to the mobility of doctoral candidates and postdoctoral fellows from these countries to the U.S. or U.K. or perhaps more generally to Western Europe.

The semantic map (Fig. 2) suggests the thematic structure of the major concepts of the emerging field in terms of the (1) collection, (2) storage, (3) processing, and (4) analysis/visualization of data. These principal components in the data matrix reveal the cognitive structure of co-authorship at the global level. However, their coupling to the geographic network (Fig. 3) indicates an emerging division of labor in which “data” and “data mining” are related to the U.S. and also to China to some extent. On the other hand, the E.U. and Israel formed a separate cluster that was less specific in its vocabulary except for the term “gene.” The rest of the world and other concepts tended to be blended into the third group that was more heterogeneous. The general lack of connections in the U.S. and E.U. networks in terms of dominant semantics may reflect not only some differences in policy incentives across the Atlantic but also a relatively weak intellectual framework in emerging semantics. Note that central terms such as “big data” were part of the heterogeneous group.

The results suggest that a combination of words and locations in a two-mode network can provide a richer representation of the emerging field of big data science than the sum of two representations (Leydesdorff, 2010a, 2010b). This has long been known for combining words as indicators of variations and citations as those of continuity and changes along time axes (e.g., Braam et al., 1991; Lucio-Arias & Leydesdorff, 2009), but in the case of dispersed fields or those fields driven by policy initiatives, a combination of geographic and semantic perspectives may be more informative. Therefore, such a perspective can complement informetric research in addressing questions concerning the emergence and interdisciplinarity of big data research.

## Acknowledgements

The authors would like to thank Seong-Cheol Choi, Ji-Young Kim, Ji-Young Park, Yun-Cheol Heo, and Min-Ho So for their assistance during the collection and organization of data and acknowledge support from the SSK (Social Science Korea) Program of the National Research Foundation of Korea (NRF-2010-330-B00232).

## References

- Abbasi, A., Chung, K. S. K., & Hossain, L. (2012). Egocentric analysis of co-authorship network structure, position and performance. *Information Processing and Management*, 48(4), 671–679.
- Anderson, C. (2008). *The end of theory: The data deluge makes the scientific method obsolete*. WIRED. <http://www.wired.com/science/discoveries/magazine/16-07/pb.theory>
- Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004a). An international framework to promote access to data. *Science*, 303, 1777–1778.
- Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., et al. (2004b). Promoting access to public research data for scientific, economic, and social development. *Data Science Journal*, 29(3), 135–152.
- Bonacich, P. (2004). The invasion of the physicists. *Social Networks*, 26(3), 285–288.
- Borgatti, S. P. (2012). *Netdraw network visualization*. Harvard, MA: Analytic Technologies.
- Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for windows: Software for social network analysis*. Massachusetts: Analytic Technologies.
- Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis. I. Structural aspects. *Journal of the American Society for Information Science*, 42(4), 233–251.
- Burt, R. S. (1992). *Structural holes: The social structure of competition*. Cambridge, MA: Harvard University Press.
- Cho, S. E., Choi, M. G., & Park, H. W. (2012). Government-civic group conflicts and communication strategies: A text analysis of TV debates on Korea's import of U.S. beef. *Journal of Contemporary Eastern Asia*, 11(1), 1–20.
- Freeman, L. C., Borgatti, S. P., & White, D. R. (1991). Centrality in valued graphs: A measure of betweenness based on network flow. *Social Networks*, 13, 141–154.
- Freeman, L. C. (1979). Centrality in social networks conceptual clarification. *Social Networks*, 1(3), 215–239.
- Gilpress. (2012 April 26). *A very short history of Data Science*. <http://whatsthebigdata.com/2012/04/26/a-very-short-history-of-data-science/>
- Graham, M. (2012). Big data and the end of theory? *The Guardian*. <http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory>
- Halevi, G., & Moed, H. F. (2012). The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends*, (30) <http://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research-and-scientific-topic-overview-of-the-literature/>
- Hanneman, R. A., & Riddle, M. (2005). *Introduction to social network methods*. Retrieved from. <http://faculty.ucr.edu/~hanneman/>
- Harris, R. (2012). International Council for Science (ICSU) and the challenges of big data in science. *Research Trends*, 30, 11–12. <http://www.researchtrends.com/issue-30-september-2012/international-council-for-science-icsu-and-the-challenges-of-big-data-in-science/>
- Hey, T., Tansley, S., & Tolle, K. (2009). *The Fourth Paradigm. Data-intensive scientific discovery*, Microsoft.
- Jankowski, N. (2009). The Contours and Challenges of e-Research. In N. Jankowski (Ed.), *E-research: Transformations in Scholarly Practice*. London: Routledge, pp. 3–34.
- Kamada, T., & Kawai, S. (1989). An algorithm for drawing general undirected graph. *Information letters*, 31(1), 7–15.
- King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, 331(February (6018)), 719–721. <http://dx.doi.org/10.1126/science.1197872>

- Kwon, K.-S., Park, H. W., So, M. H., & Leydesdorff, L. (2012). Has globalization strengthened South Korea's National Research System? National and international dynamics of the triple helix of scientific co-authorship relationships in South Korea. *Scientometrics*, 90(1), 163–176.
- Leydesdorff, L., Wagner, C., Park, H. W., & Adams, J. (2013). International collaboration in science: The global map and the network. *El profesional de la información*, 22(1), 87–94.
- Leydesdorff, L., & Bornmann, L. (2012). Testing differences statistically with the leiden ranking. *Scientometrics*, 92(3), 781–783.
- Leydesdorff, L. (2007). "Betweenness centrality" as an indicator of the "interdisciplinarity" of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303–1309.
- Leydesdorff, L. (2010a). Eugene Garfield and algorithmic historiography: Co-words, co-authors, and journal names. *Annals of Library and Information Studies*, 57(3), 248–260.
- Leydesdorff, L. (2010b). What can heterogeneity add to the scientometric map? Steps towards algorithmic historiography. In M. Akrich, Y. Barthe, F. Muniesa, & P. Mustar (Eds.), *Débordements: Mélanges offerts à Michel Callon* (pp. 283–289). Paris: École Nationale Supérieure des Mines, Presses des Mines.
- Loukides, M. (2012). *What is data science? The future belongs to the companies and people that turn data into products*. Sebastopol, CA: O'Reilly Media. <http://radar.oreilly.com/2010/06/what-is-data-science.html#data-scientists>
- Lucio-Arias, D., & Leydesdorff, L. (2009). The dynamics of exchanges and references among scientific texts, and the autopoiesis of discursive knowledge. *Journal of Informetrics*, 3(2), 261–271.
- McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company. <http://www.mckinsey.com/mgi>
- Miller, H. E. (2013). Big-data in cloud computing: A taxonomy of risks. *Information Research*, 18(1). Available at <http://InformationR.net/ir/18-1/paper571.html> (Paper 571)
- Naur, P. (1974). *Concise Survey of Computer Methods*. Lund, Sweden: Studentlitteratur AB.
- OECD. (2012 October 16). *OECD Technology Foresight Forum 2012 – Harnessing data as a new source of growth: Big data analytics and policies*. Paris, France: OECD Headquarters.
- O'Reilly Radar Team. (2012). *Planning for big data: A CIO's handbook to the changing data landscape*. Sebastopol, CA: O'Reilly Media.
- Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, 28, 441–453.
- Park, H. W. (2012). How do social scientists use link data from search engines to understand Internet-based political and electoral communication. *Quality and Quantity*, 46(2), 679–693.
- Park, H. W., & Leydesdorff, L. (2010). Longitudinal trends in networks of university–industry–government relations in South Korea: The role of programmatic incentives. *Research Policy*, 39(5), 640–649.
- Park, H. W., & Leydesdorff, L. (2008). Korean journals in the Science Citation Index: What do they reveal about the intellectual structure of S&T in Korea? *Scientometrics*, 75(3), 439–462.
- Rogers, S. (2011). *Data journalism at the guardian: What is it and how do we do it?* <http://www.guardian.co.uk/news/datablog/2011/jul/28/data-journalism>
- Rousseau, R. (2013, February 27). *Keyword selection for big data papers*. Personal communication.
- Rousseau, R. (2012). A view on big data and its relation to Informetrics. *Chinese Journal of Library and Information Science*, 5(3), 12–26.
- Savage, M., & Burrows, R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5), 885–899.
- Science staff. (2011). Challenges and opportunities. An introduction to special issue. *Science*, 331(6018), 692–693. <http://dx.doi.org/10.1126/science.331.6018.692> <http://www.sciencemag.org/content/331/6018/692.short>
- Shapiro, M., & Park, H. W. (2012). Regional development in South Korea: Accounting for research area in centrality and networks. *Scientometrics*, 90(1), 271–287.
- Snijders, C., Matzat, U., & Reips, U.-D. (2012). 'Big data': Big gaps of knowledge in the field of Internet science. *International Journal of Internet Science*, 7(1), 1–5.
- Wagner, C. S. (2008). *The new invisible college*. Washington, DC: Brookings Press.
- Wagner, C. S., Leydesdorff, L., & Adams, J. (2013). Policy implications of the global network of science (in preparation).
- Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.

**Han Woo Park** is an associate professor in the department of media and communication at Yeungnam University, South Korea. He is also the director of the CyberEmotions Research Institute at Yeungnam University. He was the principal investigator of the World Class University (WCU) Webometrics project (2009–2011). His research is available at <http://www.hanpark.net>.

**Loet Leydesdorff** (PhD, sociology; MA, philosophy; MSc, biochemistry) is a professor at the Amsterdam School of Communication Research (ASCoR) of the University of Amsterdam. He is a visiting professor at the Institute of Scientific and Technical Information of China (ISTIC) in Beijing and an honorary professor at the Science and Technology Policy Research Unit (SPRU) of the University of Sussex. He has published extensively in systems theory, social network analysis, scientometrics, and the sociology of innovation (for his works, see <http://www.leydesdorff.net/list.htm>).