

## Visualizing the Marrow of Science

SCImago Research Group: Félix de Moya-Anegón<sup>1</sup>, Benjamín Vargas-Quesada<sup>2</sup>, Zaida Chinchilla-Rodríguez<sup>1</sup>, Elena Corera-Álvarez<sup>1</sup>, Francisco J. Muñoz-Fernández<sup>1</sup>, Víctor Herrero-Solana<sup>1</sup>

1 University of Granada, Library and Information Science Faculty, 18071 Granada (Spain)

2 University of Alcalá, Library and Information Science Faculty, 28801 Alcalá de Henares, Madrid (Spain)

### Abstract

This study proposes a new methodology that allows for the generation of scientograms of major scientific domains, constructed on the basis of cocitation of ISI categories, and pruned using Pathfinder Network, with a layout determined by algorithms of the spring-embedder type (Kamada-Kawai), then corroborated structurally by Factor analysis. We present the complete scientogram of the world for the year 2002. It integrates the natural sciences, the social sciences, arts and humanities. Its basic structure and the essential relationships therein are revealed, allowing us to analyze at the same time the macrostructure, microstructure and marrow of worldwide scientific output.

**Keywords:** Information visualization, Domain analysis, Cocitation

### 1. Introduction

The construction of a great map of the sciences is a persistent idea of the modern ages. This need arises from the general conviction that an image or graphic representation of a domain favors and facilitates its comprehension and analysis, regardless of who is on the receiving end of the depiction, whether a newcomer or an expert. Science maps can be very useful for navigating around in scientific literature and the representation of its spatial relations (Garfield, 1986). They are optimal adequate means for representing the spatial distribution of the areas of research, while at the same time offering additional information through the possibility of contemplating these relationships (Small and Garfield, 1985). From a general viewpoint, science maps reflect the relationships between and among disciplines; but the positioning of their tags clues us into semantic connections while also serving as an index to comprehend why certain nodes or fields are connected with others. Moreover, these large scale maps of science show which special fields are most productively involved in research, providing a glimpse of changes in the panorama, and which particular individuals, publications, institutions, regions or countries are the most prominent ones (Garfield, 1994).

The construction of maps from bibliometric information is also known as scientography. According to Garfield, this term was coined by the person in charge of basic research at the

Institute of Scientific Information (ISI), George Vladutz, to denominate the graphs or maps obtained as a consequence of combining scientometrics with geography (Garfield, 1986). Although scientography not a widely familiar term, possibly due to the proliferation of terms such as “domain visualization” or “information/knowledge visualization” that make reference to similar notions, in our opinion it is the most adequate term for describing the action and effect of drawing charts of scientific output.

And so scientography, by means of its product known as scientograms, has become a tool and method for the analysis of domains in the sense used by Hjørland and Albrechtsen (1995), consolidating the holistic and realistic focuses of this type of analysis. It is a tool in that it allows the generation of maps; and a method in that it facilitates the analysis of domains, by showing the structure and relations of the inherent elements represented. In a nutshell, scientography is a holistic tool for expressing the discourse of the scientific community it aspires to represent, reflecting with accuracy the intellectual consensus of researchers making up that community, on the basis of their own citations of scientific literature.

The present submittal is the first of three works in which we approach the visualization, analysis, comparison and evolution of vast scientific domains. Here, we put forth a new methodology for the visualization and analysis of the greatest scientific domain imaginable: the world. A second work will expound the methodology for a comparative analysis of the major geographical and scientific domains of the US and the European Union. In a third contribution (still under construction) we shall propose a methodology for the dynamic analysis of these same domains.

First, we offer a brief overview of all work to date related with our proposal. We proceed then to outline the methodological development and its validation. After showing the results obtained, an analysis is offered on three levels — the macrostructure, the microstructure, and the marrow of recent scientific output. The ensuing discussion leads us to some brief final conclusions.

## **2. Related works**

In Moya Anegón [et al.], (2004), we ventured forth with a historic evolution of scientific maps from their origin up to the present, and proposed ISI - JCR category cocitation for the representation of major scientific domains. Its utility was demonstrated by a visualization of the scientific domain of geographical Spain for the year 2000. Since then, other works related with the visualization of great scientific domains have appeared. However, all use journals as the unit of analysis, with the exception of a study based on the cocitation of categories (Moya Anegón [et al.], 2005), comparatively focusing on the geographic domains of three countries (England, France and Spain). In contrast, Leydesdorff, (2004a); (2004b) classifies world science using the graph - analytical algorithm of bi-connected components in combination with JCR 2001. Boyack; Klavans; and Börner, (2005), apply eight alternative measures of journal similarity to a data set of 7,121 journals covering over one million documents in the combined Science Citation and Social Science Citation Indexes, to show the first global map of science using the force - directed graph layout tool VxOrd. Samoylenko [et al.](in press)propose an approach to view the scientific world and its evolution by constructing minimum spanning trees of scientific journals using the Science Citation Index from 1994 to 2001.

In our particular attempt to visualize major scientific domains, we propose the generation of scientograms through the cocitation of ISI-JCR categories, pruned by means of PathfinderNetworks (PFNET) layout (Schvaneveldt, 1990), using the algorithm of Kamada and Kawai, (1989). This is applied to world scientific output as computed by the databases Science Citation Index - Expanded (SCI-EXPANDED), Social Science Citation Index (SSCI) and the Arts and Humanities Citation Index (A&HCI).

### **3. Methodology**

In processing and depicting the scientific structure of great domains, we further developed a methodology that follows the flow of knowledge domains and their mapping as proposed by Börner; Chen; and Boyack, (2003).

#### ***3.1 Data source and processing***

Although there may be a number of alternative starting points for such an objective, — for example: Scopus (Elsevier B. V., 2005) — and points such as the bias in territorial coverage, idiomatic restrictions and the documental typology must be acknowledged, we believe that at present the ISI databases still reflect in a consistent and quite accurate way, world research having international visibility. Proof of this lies in the fact that they are used the world over for formally evaluating research activity. With this understanding, and for strictly investigative purposes, on the 2<sup>nd</sup> of August of 2004 we finished downloading from the Web of Science (The Thomson Corporation, 2005b) — and more specifically from the SCI - EXPANDED, SSCI and A&HCI — all the records of world scientific production published in the year 2002 (all those that in the field “Year” contained the string of characters corresponding to 2002). Because ISI assigns each journal to one or more subject categories, in order to designate a subject matter (ISI category) for each document, we also downloaded the Journal Citation Report (JCR) (The Thomson Corporation, 2005a), in both its Science and Social Sciences editions, for 2002. The downloaded records were exported to a relational database that reflects the structured information of the documents. This new repository contained nearly a million (901,493) source documents (articles, biographical items, book reviews, corrections, editorial materials, letters, meeting abstracts, news items and reviews) that had been published in 7,585 ISI journals (5,876 + 1,709) classified in a total of 219 categories, and altogether citing 25,682,754 published documents. The information was processed with a PC with a speed of 3 Mhz, 512 Mb RAM, and 120 Mb of hard disc.

#### ***3.2 Units of Measure***

The items of measure used most commonly for the representation of scientific domains are: journals, documents, authors, terms and words. Yet recently an addition was made to this list, with some broader units such as countries, subject spheres of different levels, institutions, and ISI categories.

One aspect to bear in mind with regard to units of analysis is the amount of information available, and the subsequent grand total if the entire domain to be represented. If the number of variables or items to be handled is much reduced, it is possible to build visualizations of domains with very small units such as words or descriptors. If this is not the

case, we must use broader units of analysis, as would be documents or authors. Yet if the amount of information processed is truly very high, it will be necessary to resort to units of analysis capable of containing smaller units, as is the case of journals that group documents, authors and terms; or of categories, which embrace all the above. This consideration is not new in the field of information visualization, and it stems from the physical limitations implicit in representing vast quantities of information in a reduced space. Authors such as Tufte, (1994); (2001) have analyzed different approaches adopted in the face of this and similar problems encountered in the graphic representation of information.

In our view, despite the specified drawbacks, the ISI categories are effective in classifying documental contents in their databases. As informational units they are, in themselves, sufficiently explicit so as to be used in the representation of all the disciplines that make up science in general. These categories, in combination with the adequate techniques for the reduction of space and the representation of the information to construct scientograms of science or of major scientific domains, prove much more informative and user friendly for quick comprehension and handling by non-expert users than those obtained by the cocitation of smaller units of cocitation. The latter would require tagging—usually involving human intervention—of the clusters generated in order to make for a comprehensible representation.

For these reasons, we used the 219 categories of the JCR 2002 as units of measure, with the exception of Multidisciplinary Sciences. One of the characteristics of JCR is that it assigns this category to a specific group of journals of a multidisciplinary nature such as Science, Nature, Endeavor or PNAS, among others. While this may seem logical and accurate at first, if we stop to look more closely at the consequences of this decision we see that works dealing with a given discipline such as Genetics appear cut off from akin categories, simply by virtue of having been published in a multidisciplinary journal and tagged only as such. This problem is not easy to solve without the human touch. The group at the University of Leyden, in the Third European Report on Scientific and Technological Indicators (European Commission, 2003), directly eliminates the category Multidisciplinary Sciences and therefore does not include the documents under this heading. We preferred to avoid losses of information and assigned, instead, a specific category to each of the documents published in journals of Multidisciplinary Sciences. This new category would be the one most referenced by each pertinent document. In the case where the category “Multidisciplinary sciences” was the only one referenced for that document, we retrieved the source documents from within and, judging from the title and abstract, manually assigned it a JCR category (one or more). The maximum number of categories with which we worked, then, was 218. The categorization of all the database documents by means of this procedure reduced the number of errors of document categorization that would have no doubt resulted from the mechanical assignment of these documents to journal subject categories.

### ***3.3 Similarity measures***

In light of our previous experience (Moya Anegón [et al.], 2004; Moya Anegón [et al.], 2005), we use cocitation as the similarity measure to quantify the relationship existing between each one of the JCR categories. We have seen in the past that the introduction of measures of

standardization in the values of cocitation matrixes, whether using that of Pearson, the cosine function (Salton; Allan; and Buckley, 1994) or Salton and Bergmark's, (1979) measure of cocitation normalization, all cause distortions in the visualization of the information, as recently described by Leydesdorff and Vaughan, (In press). Therefore, after a number of trials, we arrived at the conclusion that using tools of Network Analysis, the best visualizations are those obtained through raw data cocitation as the unit of measure.

Nonetheless, and because it is necessary to reduce the number of coincident cocitations to enhance yield of the pruning algorithm, to those raw data values we added the standardized cocitation value. In this way, we achieved a means of working with raw data cocitation, while at the same time differentiating the similarity values between categories with equal cocitation frequencies. We did this with a simple modification of the equation for the standardization of the degree of citation, proposed by Salton and Bergmark:

$$CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i) \cdot c(j)}}$$

Equation 1. Modified Measure of Cocitation, after Salton and Bergmark.

Where:

- CM is Cocitation Measure
- Cc is cocitation frequency
- c is citation
- i,j, are categories.

The result is a symmetric matrix of N by N categories, where N is the number of categories existing in the output of a domain to be visualized. These cocitation matrixes are the base and the origin of the scientograms, which show the structure of the domain represented, as well as the relationships and flows of information (knowledge) within it, between and among disciplines.

### **3.4. Layout**

#### **3.4.1. Dimensionality reduction**

Over the history of the visualization of scientific information, very different techniques have been used to reduce n-dimensional space. Either alone or in conjunction with others, the most common are: Multidimensional Scaling (MDS), Clustering, Factor Analysis (FA), Self - Organizing Maps (SOM) and PFNET.

Representing the structure of the scientific output of large domains on a plane is no easy task, regardless of whether the domain to be visualized is a region, a state, a country small or large, a continent, or even the world. The adoption of the ISI categories as units of measure implies that the resulting scientograms normally contain over 200 categories in the case of generalistic domains. To display a domain involving such a high number of units that can be easily

identified by tags, that show its interactions by means of links, and all this in an intelligible and esthetically pleasing form for the human eye... a most formidable challenge. Bearing in mind the precautionary message of Hjørland and Albrechtsen, (1995):

*If users are provided with a system of too many possibilities, without giving priority to the essential connections, the user is overloaded, and the system is ineffective.*

There is also the advice of Small, (2000):

*Despite the loss of structural information... the gain in simplicity may for some purposes be worth the sacrifice.*

And we fully agree with White, (2003) in that:

*Among techniques, two dimensional PFNET made with raw cocitation counts, and visualized through spring embedders, appears to have considerable advantages.*

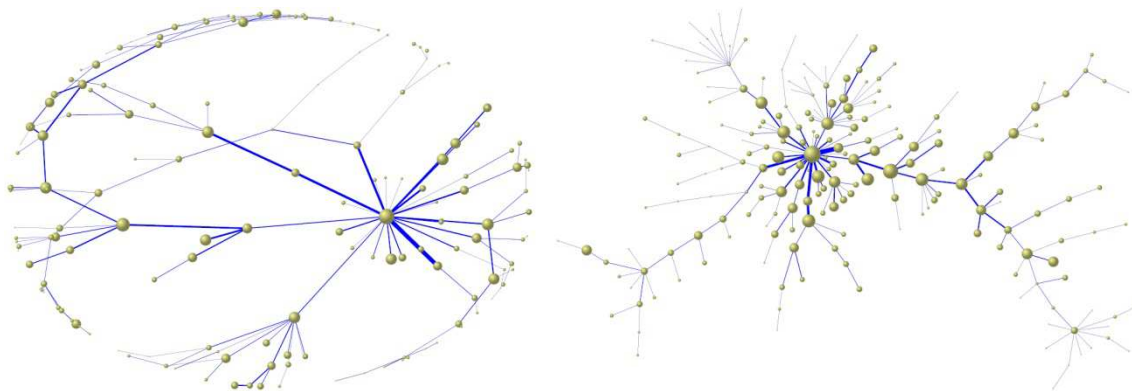
This process of schematizing information is not new to this decade. It is a principle formulated in the Middle Ages, known as the principle of simplicity, or Ockham's razor. In its original 15<sup>th</sup> century formulation, this principle was expressed in Latin: *pluralitas non est ponenda sine necessitate* ("plurality should not be posited without necessity"), which in common language could be stated as "adopt the simplest hypothesis that may explain the observations". This law of parsimony has often been taken further, especially since the laws of physics began to be expressed in the language of mathematics: The simplest hypothesis is, a priori, that which has the simplest mathematical formulation. In the case of visualization and analysis of the scientific structure, we apply this same minimalistic principle. Why visualize and analyze a dense and complex structure if we can obtain for study a simpler one, containing the most significant or essential relationships? The methodological challenge presented in these terms is, at any rate, that of the proper choice of significant links. To complete this task with rigor, the option of choice was the PFNET algorithm.

As a consequence of the interdisciplinarity of science, the matrixes proceeding from ISI subject category cocitation analysis tend to have highly interrelated elements... to such a point that the graphic representation shows a bramble of connections so tangled that they cannot be studied. In our opinion, PFNET with pruning parameters  $r = \infty$ ,  $\gamma q = n - 1$ , is the prime option for eliminating less significant relationships, while preserving and highlighting the most essential ones, and capturing the underlying intellectual structure in a economical way. Although PFNET has been used in the fields of Bibliometrics, Informetrics and Scientometrics (BIS) since 1990 (Fowler and Dearhold, 1990), its introduction in citation was the work of Chen, (1998); (1999), who introduced a new form of organizing, visualizing and accessing information. The effect is to prune all paths except those with the single highest (or tied highest) cocitation counts (White, 2001) between categories.

### 3.4.2 Scalar

There are many different methods for the automatic generation of graphs. The spring embedder type is most widely used in the area of Documentation, and specifically in domain visualization. Spring embedders begin by assigning coordinates to the nodes in such a way that the final graph will be pleasing to the eye — a process known as embedding (Eades, 1984). Two major extensions to the algorithm proposed by Eades have been developed by Kamada and Kawai, (1989) and Fruchterman and Reingold, (1991). The criteria for evaluating this type of algorithm are, basically, of an aesthetic nature: the uniform distribution of nodes, the uniform length of the links, the avoidance of crossed links, etc., all play a fundamental role in deciding to use one algorithm or another. While studies such as that of Brandenburg; Himsolt; and Rohrer, (1995) do not detect any single predominating algorithm, most of the scientific community goes with the Kamada-Kawai algorithm. The reasons upheld are its behaviour in the case of local minima, its capacity to minimize differences with respect to theoretical distances in the entire graph, its good computation times, or because it subsumes multidimensional scaling when the technique of Kruskal and Wish, (1978) is applied. As Cohen, (1997) and Krempel, (1999) indicate, the Kamada-Kawai algorithm uses a criteria or energy similar to the stress of multidimensional scaling as the measure for adaptation to theoretical distances.

We tried out hundreds of representations using the Kamada-Kawai algorithm, and compared results to those obtained with the Fruchterman and Reingold algorithm. The images shown in Figure 1 make quite clear why we preferred the Kamada-Kawai option.



**Figure 1.** Scientograms obtained using the algorithms of Fruchterman & Reingold, and Kamada & Kawai, respectively.

The result obtained by combining PFNET with the Kamada-Kawai algorithm is as spectacular and visually informative as the map of an underground metro or railroad system.

- At one glance the center and the outer limits of the system (domain) can be seen
- It is easy to get from one station (category) to another, following the trails or links

- We can determine effortlessly which are the most important nodes in terms of the number of connections received and, in turn, which points act as intermediaries with other lines, as hubs or forking points.

### **3.4 Display**

There is no clear consensus as to format on the part of researchers dedicated to domain visualization; rather, a wide variety of formats may be used, including GIF, JPG, Postscript, (PS), Encapsulated Postscript (EPS), Virtual Reality Modelling Language (VRML), or Scalable Vector Graphics (SVG), among others. In most cases, selection is conditioned by the output format of the programs used by researchers themselves. Still, it is important to obtain quality images with a low weight in bits, so that they can be easily transported over the web, in light of the growing interest surrounding vectorial graphics and their animation. We find SVG great to work with. Its format is light, quick, ingenious and... free! Being vectorial, it allows one to zoom in and out, and wander up, down or sideways, without diminishing the quality of the graphics, while at the same time allowing searches for textual information from within points in the image. Moreover, it is backed up by firm technical assistance and important underpinnings in the sector on the whole and by its individual experts. Also, being part of the XML standard permits its use as an interface by facilitating the integration of codes to control the interaction of the graphs and the user. For all these reasons it was our overriding choice as the visualization format for the scientograms of great scientific domains.

In order that the scientograms could be displayed in the vectorial format, we exported them to an SVG format by means of ad hoc software. This software was in charge of carrying out the aesthetic and informational touch-ups in the scientograms. The tasks it performs are:

- Detection of the superposition of nodes or links, so that they can be repositioned manually
- Painting the nodes with previously defined colors
- Tagging each node with the corresponding ISI category
- Insertion of hyperlinks in nodes and links, to permit the retrieval of related bibliographic information from the relational database (see Moya Anegón [et al.] (2005) for further reference).

As a result, we can create visualizations of large scientific domains that facilitate browsing throughout their structures and that favor access and retrieval of bibliographic information from behind the depiction.

### **4. Materialization and validation of the scientograms.**

The scientogram of Figure 2 shows the structural image of world scientific output in terms of ISI categories for the year 2002. It resembles a human neuron with a huge axon or central neurite. This scientogram is the visualization obtained as a consequence of applying our methodology to nearly a million scientific documents gathered in the ISI databases, which were then grouped into categories of production, for the year 2002. The links show the most relevant interactions produced between or among categories, and reflect the majority



viewpoint of some two million scientific authors in light of their early 26 million references to other work.

To enhance user comprehension of the scientogram, each sphere is tagged with the name of the JCR category that represents it, and is given a size proportional to the number of documents constituting it in the year 2002. To help visually establish the relationship between the size of each category and its true output, in the lower left part of the scientogram there is a sphere of reference — a figure scale — with a size equivalent to one thousand documents. The lines that connect the different spheres are the most significant relationships of cocitation among the categories, the least essential ones having been eliminated with PFNET. As the physical distance between each pair of adjacent categories on the map tends to be constant, the ties are thicker or thinner depending on the intensity of cocitation (the higher the cocitation, the greater the thickness).

The spatial distribution of the categories in the scientogram is determined by the tandem of raw data cocitation and PFNET. Those categories with a greater number of links (higher degree of cocitation) appear in the center. As this number diminishes, the nodes approach the periphery. Just as White, (2003) did, though with a greater number of units of measure in our case, we observe that around the most prominent categories, reminiscent of bunches of grapes, we see the great thematic areas that make up the domain, chained together in explicit sequences. The order the categories occupy in that chain is by no means arbitrary, reflecting how the subject areas are connected amongst themselves. In this way, the substructures generated from the prominent subject categories reveal the major thematic areas, while the connections among prominent categories reveal how these are interrelated. For example, a second look at Figure 2 allows us to distinguish a huge central cluster surrounded by other smaller ones, distributed all over the surface of the scientogram. If we look even more closely at this central bunch and then at another lower one, we will discover the following chain: Biochemistry & Molecular Biology  $\leftrightarrow$  Neurosciences  $\leftrightarrow$  Clinical Neurology  $\leftrightarrow$  Psychiatry  $\leftrightarrow$  Psychology. This path indicates that in the scientogram there are two major subject areas that we could denominate Biomedicine and Psychology, whose most prominent categories, respectively, are Biochemistry & Molecular Biology and Psychology; which in turn are connected by intermediary categories such as Neurosciences, Clinical Neurology, and Psychiatry. The same can be said, for example, of the chain in the left midsection: Mathematics Miscellaneous  $\leftrightarrow$  Social Sciences  $\leftrightarrow$  Mathematical Methods  $\leftrightarrow$  Economics  $\leftrightarrow$  History of Social Sciences  $\leftrightarrow$  History, which shows how Mathematics is indirectly connected with Humanities. These paths are very important, as they are perceived as the thread uniting the overall scientific structure of a domain.



We stopped extracting factors upon arriving at an eigenvalue greater or equal to one<sup>1</sup>, which was done with the Scree test<sup>2</sup>. In order to capture the nature of each factor so as to tag it, we followed the methodology proposed by Moya Anegón; Jiménez Contreras; and Moneda Carrochano, (1998). The factors were first ordered according to their index of weight — factor loading — in a decreasing order, and a cutoff of 0.5 was established for membership; though for denomination, we took into account only those categories of each factor that had a value of 0.7 or more.

In order to set off the different subject areas while reflecting the factors in common, we gave the same color to all categories sharing a common factor, though in a range of tones. Those that belong to more than one subject area are red, the “hot” points of interaction among the subject areas. Finally, dark grey shows the “cold” ones that were not identified and therefore belong to no subject area. Our findings coincide with those of Boyack; Klavans; and Börner, (2005) in that certain categories are not adequately represented by the documents that make them up, for instance in Mathematics. This is due to the fact that some journals publish articles whose contents have very little to do with the ISI tag per se, which is not a problem exclusive to categories, but may also stem from the journals themselves.

FA captures quite effectively the structure of PFNET and viceversa, which comes as no surprise, since FA works with the categories with the highest loading factor, whereas PFNET takes those with the greatest degree of cocitation, and maintains the nodes of the network that are interconnected. FA is good at identifying within the scientogram the consolidated thematic areas, or those consisting of a category with a high degree of interconnection or prominence. On the other hand, it has trouble detecting the least relevant categories, and therefore the least consolidated corresponding subject areas; it may even identify special fields within a single subject area as full - fledged categories. PFNET, with its bunches, illustrates the possible subject areas of a scientogram, identifies the most prominent category of each, and indicates the route of connection of certain areas with other ones, by means of the sequencing of categories that connect them. However it does not provide any information regarding the limits of the given subject area, nor does it give clues as to its denomination. We therefore hold PFNET and FA to be extremely valuable as complements in the detection of the structure of a scientific domain, the advantages of one making up for the drawbacks of the other in a mutual way. Thus, FA is responsible for identifying, delimiting and denominating the great thematic areas reflected in the scientogram, though at times it mistakenly identifies particular specialized fields as subject areas. Meanwhile, PFNET is in charge of making the subject areas more visible, grouping their categories into bunches, and showing the paths that connect the different prominent categories and, finally, the overall topology of the domain.

---

<sup>1</sup> This simple criterion works quite well, giving results much in accord with the expectations of researchers (Ding; Chowdhury; and Foo, 1999).

<sup>2</sup> The scree test consists of the examination of the line obtained in the graphic representation of the eigenvalues of the identified factors. The extraction of factors comes to a halt when the line of eigenvalues begins to level out, practically forming a line parallel to the axis, with hardly any slant (Lewis-Beck, 1994).

In summary, this methodology allows for the statistical validation of the structural coherence of the scientograms of a vast scientific domain. Moreover, it brings into view the large subject areas that make up the domain, providing an image of an intellectual superstructure reminiscent of neural circuitry, which we could call a factor scientogram.

## 5. Results

### 5.1. The structure of world science

FA identifies 35 factors in the cocitation matrix of 218 x 218 categories of world science 2002. Through the scree test we extracted 16, which we tagged using the previously explained method; these accumulate 70.2% of the variance (Table 1)

Factor	Label	Eigenvalue	% variance	% cumulative
1	Biomedicine	42.255	19.4	19.4
2	Psychology	24.14	11.1	30.5
3	Material Science & Physis Applied	15.472	7.1	37.6
4	Earth & Spaces Sciences	12.655	5.8	43.4
5	Business, Law & Economy	10.069	4.6	48
6	Computer Science & Telecommunications	8.272	3.8	51.8
7	Agricultura & Soil Sciences	6.815	3.1	54.9
8	Human Studies	6.298	2.9	57.8
9	Chemistry	4.668	2.1	59.9
10	Etology	4.517	2.1	62
11	Engineering	4.195	1.9	63.9
12	Health Care & Service	3.601	1.7	65.6
13	Applied Mathematics	3.029	1.4	67
14	Nuclear Physics, Particles & Fields	2.567	1.2	68.1
15	Animal Biology & Ecology	2.321	1.1	69.2
16	Orthopedics	2.16	1	70.2

**Table 1.** Factors of the world science domain, 2002.

The number of categories included in at least one factor is 195. Twenty-three were not included in any factor (Table 2), and 25 belonged to two factors simultaneously (Table 5).

ISI Categories	
ARCHAEOLOGY	IMAGING SCIENCE & PHOTOGRAPHIC TECHNOLOGY
ARCHITECTURE	LITERATURE, SLAVIC
BUSINESS	MANAGEMENT
COMPUTER SCIENCE, INTERDISCIPLINARY APPLICATIONS	MATERIALS SCIENCE, BIOMATERIALS

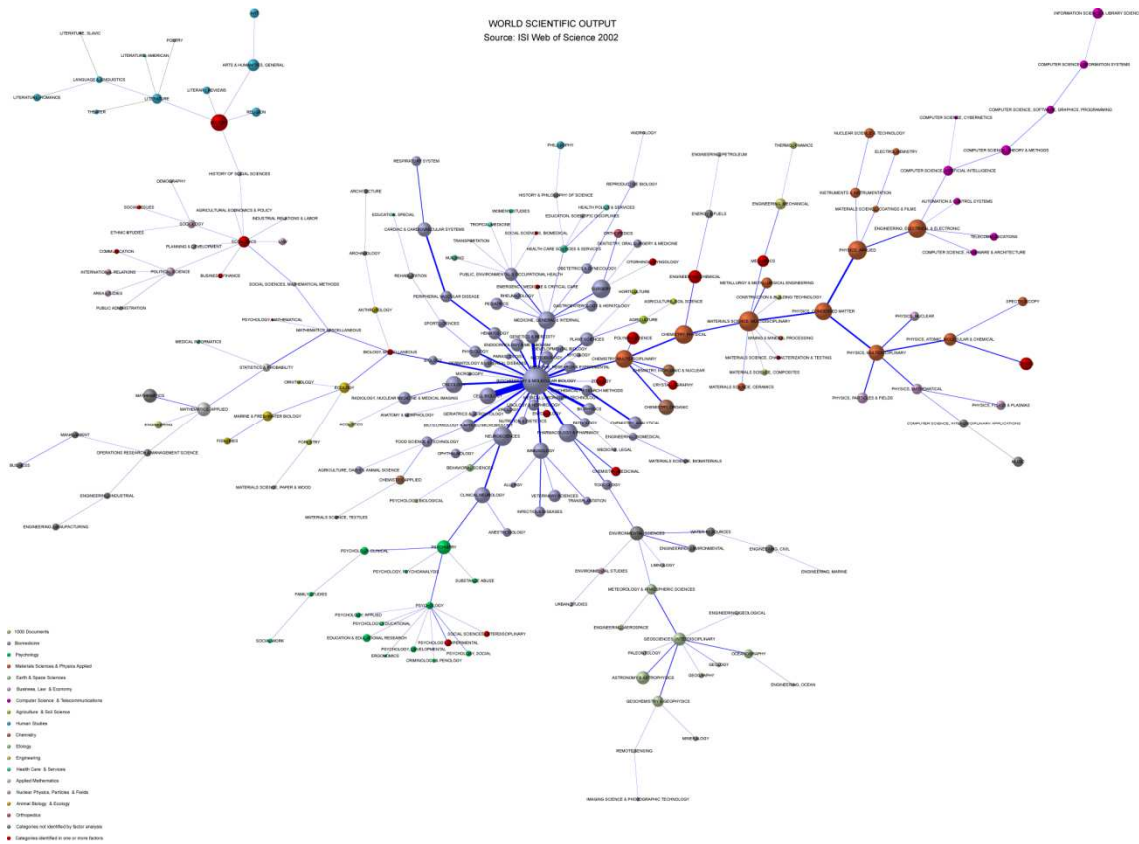
ENERGY & FUELS	MATERIALS SCIENCE, PAPER & WOOD
ENGINEERING, CIVIL	MATHEMATICS
ENGINEERING, ENVIRONMENTAL	MUSIC
ENGINEERING, INDUSTRIAL	OPERATIONS RESEARCH & MANAGEMENT SCIENCE
ENGINEERING, MANUFACTURING	REHABILITATION
ENGINEERING, MARINE	TRANSPORTATION
ENVIRONMENTAL SCIENCES	WATER RESOURCES
HISTORY & PHILOSOPHY OF SCIENCE	

**Table 2.** Categories not included in any factor.

The superposition or overlapping of these results on the scientogram give rise to the Factor Scientogram (FS) shown in Figure 3. Below, we invite the observer on an excursion over its surface. Occupying all the central area and most of the upper area, we can see in purple Biomedicine. Just above it, in deep-sea green, is Health Care & Services, and in a salmon shade, Orthopedics. Agriculture and Soil Sciences takes on a grassy green. Going clockwise, we see the rest of the thematic areas. To the center right appears Materials Science and Physics Applied, in peach. Connected to it by its top section is Engineering, in light yellow, and Computer Science & Telecommunications in hot pink. And connected to its lower part, we can see Nuclear Physics & Particles & Fields in mauve, and Chemistry in brown. The lower central zone holds Earth and Space Sciences, standing out in grey-green, and Psychology in emerald green; just above which we find Etiology, in very light green. In the left center of the display we see, in yellow, Animal Biology & Ecology. Connected to it is Applied Mathematics in dark grey, in the lower part; and in the upper part Business, Law & Economy, in light purple, and Humanities in sky blue. To establish a quick correspondence between color and the name of each thematic area, in the low left section of each FS we give a legend with the color code.

Bearing in mind that our scientograms are extremely schematic depictions of the scientific output of a domain, their analysis and interpretation will be based on inferences from the resulting PFNET structure. That is, a category or thematic area occupying a central position in the scientogram will have a more general or universal nature in the domain as a consequence of the number of sources they share with the rest, contributing more to scientific development than those with a less central position. In the same way, and generally speaking, the more peripheral the situation of a category or subject area, the more exclusive its nature, and the fewer sources it will appear to share with other categories; accordingly, the lesser its contribution to the development of knowledge through scientific publications. An intermediary position makes possible the interconnection of other categories or thematic areas. For instance, if the thematic area of Biomedicine disappeared from the scientogram of Figure 3, the rest of the areas would be left disconnected; and a similar situation would occur if Biochemistry & Molecular Biology were eliminated. We could say the same of other areas and categories, though the loss of interconnection would be less severe, as the positions involved are less central. This broad interpretation of our scientograms not only explains the patterns of

cocitation that characterize a domain; it also foments an intuitive way for specialists and non-experts to arrive at a practical explanation of the workings of PFNET (Chen and Carr, 1999).



**Figure 3.** Factor scientogram of world science, 2002.

## 5.2. Macrostructure

The first thing that strikes us when we look at the FS is the combination of just a few thematic areas that are very large in size, connected with many other small ones. This reflects the hyperbolic nature of bibliometric distributions (Small and Garfield, 1985). Another noteworthy aspect is the central-peripheral pattern that the thematic areas adopt in their manner of connection, where a large central thematic area serves as the node of connection to smaller surrounding ones. The conception of a structure formed by a center and a periphery stands as a classical paradigm and appears in many fields of science (Everett and Borgatti, 1999). The existence of a structure made up of an active nucleus, formed by a dense and compact grid of categories, creates a striking contrast with a disperse conglomerate of weak interrelations. From a macro structural point of view we can distinguish three major zones. In the center is what we could call Medical and Earth Sciences, consisting of: Biomedicine, Psychology, Etiology, Animal Biology & Ecology, Health Care & Service, Orthopedics, Earth & Space Science, and Agriculture & Soil Sciences. To the right we can see some other basic and experimental sciences: Materials Sciences & Applied Physics, Engineering, Computer Science & Telecommunications, Nuclear Physics & Particles & Fields, and Chemistry. And to the left is the neighborhood of the social sciences, with Applied Mathematics, Business, Law and Economy,

and Humanities. This scheme of macro structural vertebration of the sciences can be seen as a typical distribution in the FS displays of developed countries, but evidently differs from the scientific underpinnings of less developed nations, as we have confirmed with other FSs (Vargas-Quesada, 2005).

At a glance, the most central thematic area is Biomedicine, but to corroborate this we resort to the Social Network Analysis (Wasserman and Faust, 1998), to focus in on the degree of interconnection of the diverse thematic areas involved. This perspective leads us to reconfirm that the most central area is Biomedicine (Table 3).

Thematic Area	Grade
<i>Biomedicina</i>	8
Material Science & Physiys Applied	4
Animal Biology & Ecology	2
Business, Law & Economy	2
Applied Mathematics	2
Chemistry	2
Agricultura & Soil Sciences	1
Earth & Spaces Sciences	1
Etology	1
Nuclear Physics, Particles & Fields	1
Human Studies	1
Computer Science & Telecommunications	1
Engineering	1
Orthopedics	1
Health Care & Service	1
Psychology	1

**Table 3.** Centrality of degree of the thematic areas of world science, 2002.

The centrality of Biomedicine signals it as the area sharing more sources and contributing most knowledge to the rest, lending cohesion to the domain. The identification of Biomedicine as one of the centers of science, the thematic areas identified, their relative positions and the way they are interconnected is nearly identical to the pattern revealed by Boyack; Klavans; and Börner, (2005) in their map of the backbone of science.

The degree of universality of the rest of the thematic areas will depend on their distance from the center. The shorter this distance, the greater the involvement in domain evolution. Now

using the paths between thematic areas as the units of measure, we obtain the following ranking of universality, with Biomedicine as the point of reference (Table 4).

Thematic Area	Distance
Psychology	1
Agricultura & Soil Sciences	1
Chemistry	1
Etology	1
Health Care & Service	1
Applied Mathematics	1
Animal Biology & Ecology	1
Orthopedics	1
Material Science & Physiyys Applied	2
Business, Law & Economy	2
Earth & Spaces Sciences	3
Computer Science & Telecommunications	3
Human Studies	3
Engineering	3
Nuclear Physics, Particles & Fields	3

**Table 4.** Distances with respect to Biomedicine.

Finally, in the FS we also find a series of categories in red (suggestive of friction or points of interaction of different thematic areas). The interdisciplinary categories of the world science domain for 2002 are listed alongside the areas to which they belong in Table 5.

ISI Categories	Thematic Areas	
BIOLOGY, MISCELLANEOUS	Biomedicine	Animal Biology & Ecology
ENTOMOLOGY	Biomedicine	Animal Biology & Ecology
ZOOLOGY	Biomedicine	Animal Biology & Ecology
CHEMISTRY, MEDICINAL	Biomedicine	Chemistry
EMERGENCY MEDICINE & CRITICAL CARE	Biomedicine	Orthopedics
OTORHINOLARYNGOLOGY	Biomedicine	Orthopedics
BUSINESS, FINANCE	Business, Law & Economy	Applied Mathematics
ECONOMICS	Business, Law & Economy	Applied Mathematics



HISTORY	Business, Law & Economy	Humanities
CRYSTALLOGRAPHY	Materials Sciences & Physics Applied	Chemistry
ENGINEERING, CHEMICAL	Materials Sciences & Physics Applied	Chemistry
POLYMER SCIENCE	Materials Sciences & Physics Applied	Chemistry
MATERIALS SCIENCE, CHARACTERIZATION & TESTING	Materials Sciences & Physics Applied	Engineering
MECHANICS	Materials Sciences & Physics Applied	Engineering
OPTICS	Materials Sciences & Physics Applied	Nuclear Physics, Particles & Fields
PSYCHOLOGY, MATHEMATICAL	Psychology	Applied Mathematics
COMMUNICATION	Psychology	Business, Law & Economy
SOCIAL ISSUES	Psychology	Business, Law & Economy
SOCIAL SCIENCES, INTERDISCIPLINARY	Psychology	Business, Law & Economy
PSYCHOLOGY, EXPERIMENTAL	Psychology	Etology
SOCIAL SCIENCES, BIOMEDICAL	Psychology	Health Care & Service

**Table 5.** Categories with double thematic adscription in the World Science Domain, 2002.

### 5.3. Microstructure

The FS of Figure 3 consists of 218 categories and 217 links that interconnect them. None appears alone or disconnected. As with the thematic areas, the first thing we notice is the existence of just a few large categories, and a great number of small ones. The bigger ones are seen above all in the center and the right center of the FS, and less so to the left, meaning more production on the part of the categories under medical sciences and “hard sciences” than among the softer sciences. The pattern of connection that the categories adopt is also of the central-peripheral type: we have a large central category that functions as the central hub of the surrounding categories, while maintaining structural cohesion. There is no doubt that the most central category is Biochemistry and Molecular Biology. This is also demonstrated by its high centrality degree (Table 6).

ISI Category	Degree
BIOCHEMISTRY & MOLECULAR BIOLOGY	31
PSYCHOLOGY	10
MEDICINE, GENERAL & INTERNAL	9
MATERIALS SCIENCE, MULTIDISCIPLINARY	9

ECONOMICS	9
GEOSCIENCES, INTERDISCIPLINARY	8
CHEMISTRY, MULTIDISCIPLINARY	6
ENVIRONMENTAL SCIENCES	6
PUBLIC, ENVIRONMENTAL & OCCUPATIONAL HEALTH	6
PHYSICS, MULTIDISCIPLINARY	5
PSYCHIATRY	5
IMMUNOLOGY	5
ENGINEERING, ELECTRICAL & ELECTRONIC	5
POLITICAL SCIENCE	5
HISTORY	5
LITERATURA	5

**Table 6.** Top 16 categories of highest grade, world science 2002.

Again, Biochemistry and Molecular Biology is the category with most shared sources and the greatest share of contributions, demonstrating connectivity and intellectual interchange, emerging as a central axis of the vertebration of science in the year 2002. If we eliminated Biochemistry and Molecular Biology, the categories around it would be left disconnected, and the semantic structure of the scientogram would be dismantled. Translating the distances to the scientogram and giving each a distinctive color, we can build a new distance scientogram (DS) that visually informs, in a quick and easy manner, of the distance of each one with respect to the central category (Figure 4). A picture is worth a thousand words.

#### **5.4. The marrow of science**

The Factor Scientogram is able to reveal the marrow or essence of the worldwide productive system of science. This is achieved thanks to PFNET's capacity for selecting the most significant links among categories, together with the graphic possibilities of showing their intensity of cocitation by the thickness of the links. Going back to the FS of Figure 3, we see how there are thicker links uniting sequences of categories, highlighting the medula of this domain. To determine which links and categories constitute the spinal column of a domain, we adopt as reference the highest value of the link uniting two thematic areas, and eliminate all those that remain below this cutoff value. The result is the marrow; that is, the part of the structure of knowledge that nourishes and stimulates the whole. There are three thematic areas that stand out in the marrow scientogram (MS) of World Scientific Output 2002 (Figure 5): Biomedicine, Chemistry, and Materials Science & Physics Applied. The central position of Chemistry, in the very thick of the marrow of science, acting as a bridge between Biomedicine and Materials Science & Applied Physics, coincides with the findings of Valente and Hopkins, (2005).

Noteworthy is the fact that, despite the extreme simplification undertaken, there are still paths of long distance such as: *Cardiac & Cardiovascular Systems*  $\leftrightarrow$  *Peripheral Vascular Disease*  $\leftrightarrow$  *Hematology*  $\leftrightarrow$  *Biochemistry & Molecular Biology*, indicating the high degree of interdisciplinarity of these categories. The thematic area Materials Sciences & Physics

Applied appears as a reduced version of its very same structure in the FS display. In spite of the schematization, the medula serves to demonstrate the sequence of its basic structure, which runs from Chemistry Multidisciplinary to Physics Condensed Matter, this in turn serving as a bridge over to Physics Applied, and to Physics Atomic Molecular & Chemical.

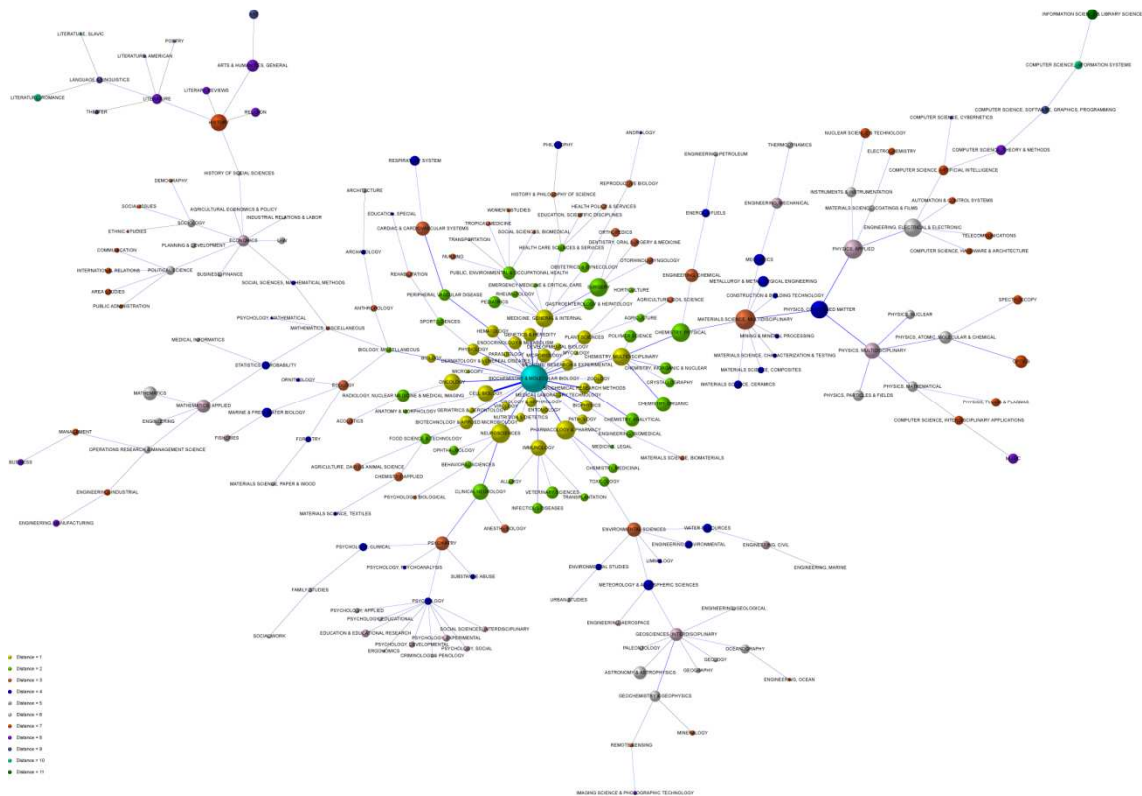


Figure 4. Scientogram of world scientific distances, 2002, with respect to the central category.

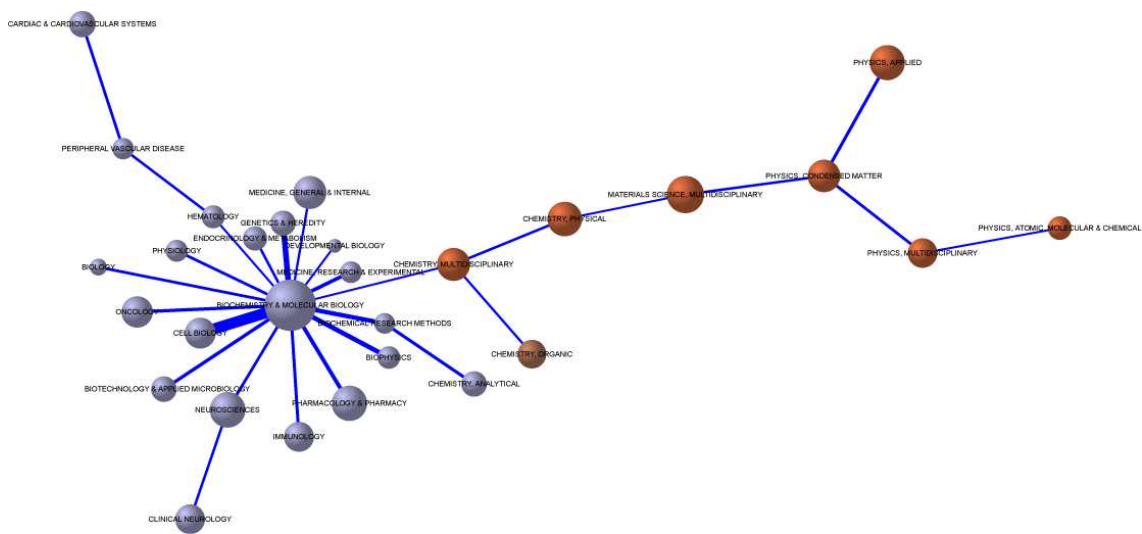


Figure 5. The marrow of world scientific output, 2002.

## 6. Discussion

The scientography of vast scientific domains wields the possibility of exploring the state of research from a wide array of perspectives. On the one hand it offers domain analysts the possibility of seeing and knowing which are the essential connections between categories of a given domain. On the other hand, it allows us to see how these categories are grouped in major thematic areas and how they are interrelated in a logical order of explicit sequences.

Scientograms are a well-designed means of domain visualization in that they can depict small or large amounts of information, they reduce the time of visual searching, they facilitate the comprehension of complex data structures, they make manifest relationships among elements that would otherwise not be acknowledged, and favor the formulation of hypotheses. As the visual result of the consensual opinion of a domain's authors (Vargas-Quesada, B. 2005) scientograms are, moreover, evidence of the evolution of science. True, they cannot predict the future horizons of research, though they may give some clues, as the changes within over a period of time can reveal tendencies that can be extrapolated to put forth a prognosis of the domain. Their topology, representing the structure of the scientific achievements of a specific time period, can be viewed sequentially or dynamically to study the evolution of a domain (<http://atlasofscience.net/spanish-evolution.svg>).

Scientograms offer new investigators, and those approaching them for the first time, a lasting image of the essential structures of a domain, which can help them complete the mental image already harboured, or become the new point of reference from which an individual perception of the scientific domain can be constructed.

## 7. Conclusions.

Our sources to date allow us to affirm that this is the first time in the history of information visualization that the world science map has been displayed and analysed in complete fashion (SCI, SSCI and A&HCI). This new methodology for the visualization and analysis of large scientific domains stands as the practical connection of several fields of research, including information visualization, citation analysis, social network analysis, and domain analysis. With very basic means and minimal informational costs, the methodology has allowed us to convey the schematic relationships existing among millions of documents and generate the complete visualization of the greatest scientific domain feasible: the world.

We consider this a very powerful tool, not only in view of its capacity to schematize, but also because of its facility for representing relational information chained in a series of intelligible sequences, which facilitate and favour comprehension, analysis and interpretation of the structure of a domain, both for neophytes and for experts. The advantages of scientograms of vast domains are many from the viewpoint of information visualization and analysis. Yet they also entail two aspects that call for improvement. The first of these is related with the information used to build the scientograms.

Although ISI databases are a most prestigious and adequate means for representing the scientific structure of any domain, the exhaustivity of scientograms would benefit from the incorporation of information proceeding from other sources, such as specialized databases and conference reports. The second area for improvement surrounds interpretation: although scientograms are the social and holistic reflection of a domain, the final interpretation is the activity of an individual who is not exempt from some degree of subjectivity. For this reason, we consider it important to continue work on the design of techniques that would enhance the objective components of information representation and limit the more subjective elements.

**Acknowledgements:** this work was made possible by public financing through the project PN I+D+I: SEJ2004-08358-CO2-01/SOC.

## References

1. Boyack, K.W., Klavans, R., and Börner, K. (2005). Mapping the backbone of science. *Scientometrics* 64, 351-374.
2. Brandenburg, F.J., Himsolt, M., and Rohrer, C. (1995). An Experimental Comparison of Force-Directed and Randomized Graph Drawing Algorithms. *Lecture Notes in Computer Science* 1027, 87.
3. Börner, K., Chen, C., and Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science & Technology*, 37, 179-255.
4. Chen, C. (1998). Bridging the gap: the use of pathfinder networks in visual navigation. *Journal of Visual Languages and Computing* 9, 267-286.
5. Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing & Management* 35, 401-420.
6. Chen, C. and Carr, L. (1999). Trailblazing the literature of hypertext: an author cocitation analysis (1989-1998). Proceeding of the 10th ACM Conference on Hypertext (Hypertext '99).
7. Cohen, J. (1997). Drawing Graphs to Convey Proximity: An Incremental Arrangement Method. *ACM Transactions on Computer-Human Interaction* 4, 197-229.
8. Ding, Y., Chowdhury, G.G., and Foo, S. (1999). Mapping the intellectual structure of information retrieval studies: an author co-citation analysis, 1987-1997. *Journal of Information Science* 25, 67-78.
9. Eades, P. (1984). A Heuristic for Graph Drawing. *Congressus Numerantium* 42, 149-160.
10. Elsevier B. V. (2005). *Scopus*. Available from: <<http://www.scopus.com>> (Visited: 17/8/2005).
11. European Commission (2003). Third European Report on Science & Technology Indicators, 2003: towards a knowledge-based economy. Brussels: Directorate-General for Research.
12. Everett, M. G. and Borgatti, S. P. (1999). Peripheries of Cohesive Subsets. *Social Networks*, 21, (4), 397-407.
13. Fowler, R. H. and Dearhold, D. W. (1990). Information retrieval using path finder networks. In R. W. Ed. Schvaneveldt (Ed.), *Pathfinder associative networks; studies in knowledge organization*. Ablex (NJ): Norwood.
14. Fruchterman, T. and Reingold, E. (1991). Graph Drawing by Force-Directed Placement. *SoftwarePractice and Experience* 21, 1129-1164.
15. Garfield, E. (1986). Towards scientography. *Essays of an Information Scientist* 9, 324.

16. Garfield, E. (1994) Scientography: mapping the tracks of science. *Current contents: Social & Behavioral Sciences* 7, 5-10.
17. Hjørland, B. and Albrechtsen, H. (1995). Toward a new horizon in information science: domain analysis. *Journal of the American Society for Information Science (JASIS)* 46, 400-425.
18. Kamada, T. and Kawai, S. (1989). An algorithm for drawing general undirected graphs. *Information Processing Letters* 31, 7-15.
19. Krempel, L. (1999). Visualizing Networks with Spring Embedders: Two-mode and Valued Graphs. International Sunbelt Social Network Conference.
20. Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills: Sage.
21. Lewis-Beck, M.S. (1994). *Factor analysis and related techniques*. London: Sage.
22. Leydesdorff, L. (2004a). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. *Journal of Documentation* 60, 371-427.
23. Leydesdorff, L. (2004b). Top-down decomposition of the journal citation report of the social science citation index: graph and factor-analytical approaches. *Scientometrics* 60, 159-180.
24. Leydesdorff, L. and Vaughan, L. (In press). Co-occurrence matrices and their applications in information science: extending ACA to the web environment. *Journal of the American Society for Information Science and Technology (JASIST)*.
25. Moya Anegón, F. de, Vargas-Quesada, B., Chinchilla-Rodríguez, Z., Herrero-Solana, V., Corera-Álvarez, E., and Muñoz-Fernández, F.J. (2005). Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. *Information Processing & Management* 41, 1520-1533.
26. Moya Anegón, F. de, Jiménez Contreras, E., and Moneda Carrochano, M.d.l. (1998). Research fronts in library and information science in Spain (1985-1994). *Scientometrics* 42, 229-246.
27. Moya Anegón, F. de [et al.] (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics* 61(1), 129-145.
28. Salton, G., Allan, J., and Buckley, C. (1994). Automatic structuring and retrieval of large text file. *Communications of the ACM* 37, 97-108.
29. Salton, G. and Bergmark, D. (1979). A citation study of computer science literature. *Professional Communication, IEEE Transaction* PC-22, 146-158.
30. Samoylenko, I., Chao, T.-C., Liu, W.-C., and Chen, C.-M. (In press) Visualizing the scientific world and its evolution. *Journal of the American Society for Information Science and Technology*.
31. Schvaneveldt, R.W. (1990). *Pathfinder Associative Networks*. Norwood, NJ: Ablex.

32. Small, H. (2000). Charting pathways through science: exploring Garfield's vision of a unified index to science. In B. Cronin and H. B. Atkins (Eds.), *The web of knowledge: a Festschrift in honor of Eugene Garfield*. Medford, (N. J.): Information Today. P. 449-473.
33. Small, H. and Garfield, E. (1985) The geography of science: disciplinary and national mappings. *Journal of Information Science* 11, 147-159.
34. The Thomson Corporation. (2005a). *ISI Journal Citation Reports*. Available from: <<http://go5.isiknowledge.com/portal.cgi>> (Visited: 3/9/2005a).
35. The Thomson Corporation. (2005b). *ISI Web of Science*. Available from: <<http://go5.isiknowledge.com/portal.cgi>> (Visited: 3/9/2005b).
36. Tufte, E.R. (1994). *Envisioning information*. Cheshire (U.K.): Graphics Press.
37. Tufte, E.R. (2001). *The visual display of quantitative information*. Cheshire (U.K.): Graphics Press.
38. Valente, T.W. and Hopkins, J. (2005). El juego de la red de difusión. *REDES – Revista hispana para el análisis de redes sociales* 8, 1-17.
39. Vargas-Quesada, B. (2005). *Visualización y análisis de grandes dominios científicos mediante redes pathfinder*. Tesis doctoral. Granada: Universidad.
40. Wasserman, S. and Faust, K. (1998). *Social network analysis: methods and applications*. Cambridge: Cambridge University Press.
41. White, H.D. (2001). Author-centered bibliometrics through CAMEOs: characterizations automatically made and edited online. *Scientometrics* 51, 607-637.
42. White, H. D. (2003). Pathfinder networks and author cocitation analysis: a remapping of paradigmatic information scientist. *Journal of the American Society for Information Science and Technology (JASIST)*, 54, (5), 423-434.