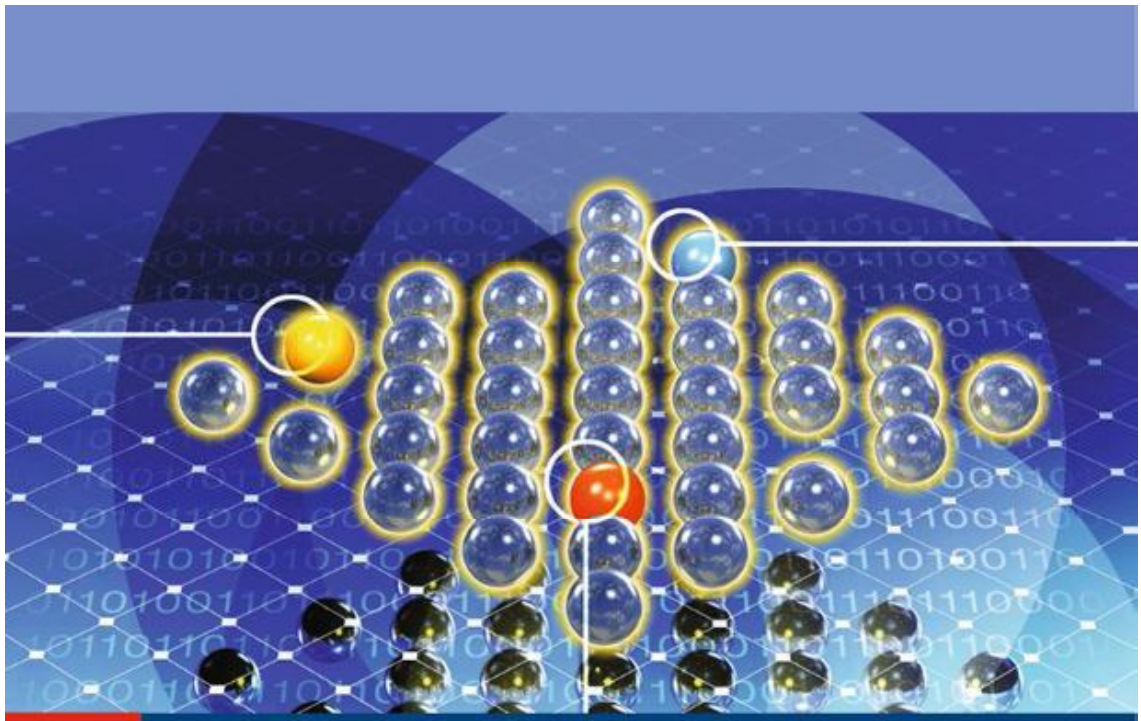


AJITH ABRAHAM
ABOUL-ELLA HASSANIEN
VÁCLAV SNÁŠEL
Editors

Computational Social Network Analysis

Trends, Tools and Research Advances

Springer

COMPUTER
COMMUNICATIONS
AND NETWORKS

Ajith Abraham • Aboul-Ella Hassanien
Václav Snášel
Editors

Computational Social Network Analysis

Trends, Tools and Research Advances

Springer

Editors
Prof. Dr. Ajith Abraham
Machine Intelligence Research
Labs (MIR)
Scientific Network for Innovation
& Research Excellence
P.O.Box 2259
Auburn WA 98071-2259
USA
ajith.abraham@ieee.org

Prof. Aboul-Ella Hassanien
Cairo University
Fac. Computers & Information
Dept. Information Technology
5 Ahmed Zewal Street
Orman, Giza 12613
Egypt
a.hassanien@fci-cu.edu.eg

Václav Snášel
Technical University Ostrava
Dept. Computer Science
Tr. 17. Listopadu 15
708 33 Ostrava
Czech Republic
vaclav.snasel@vsb.cz

Series Editor
Professor A.J. Sammes, BSc, MPhil, PhD, FBCS, CEng
Centre for Forensic Computing
Cranfield University
DCMT, Shrivenham
Swindon SN6 8LA
UK

ISSN 1617-7975
ISBN 978-1-84882-228-3 e-ISBN 978-1-84882-229-0
DOI 10.1007/978-1-84882-229-0
Springer Dordrecht Heidelberg London New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2009942893

Cover design: SPi Publisher Services

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Contents

Chapter 9
Vmap-Layout, a Layout Algorithm for Drawing Scientograms

Arnaud Quirin and Oscar Cordón

Abstract We present in this chapter a drawing algorithm to represent graphically co-citation networks (scientograms). These networks have some interesting and unusual topological properties, which are often valuable to be visualized. In general, these networks are pruned with a network scaling algorithm and then visualized using a drawing algorithm (J Vis Lang Comput 9:267–286, 1998). However, typical drawing algorithms do not work properly, especially when the size of the networks grows. Edge crossings appear while the drawing space is not adequately filled, resulting in an unsightly display. The approach presented in this chapter is able to print the networks filling all the available space in an aesthetic way, while avoiding edge crossings. The algorithm is detailed and compared with the classical Kamada–Kawai drawing algorithm on several scientograms.

9.1 Introduction

Social networks have some interesting and unusual topological properties, which are often valuable to be printed graphically. However, the raw networks cannot be often visualized easily, especially when their size grows proportionally with the number of data to be dealt with, and thus specific algorithms for simplifying such large networks have been developed. Network scaling algorithms, the goal of which is to take proximity data and to obtain structures revealing the underlying organization of those data, use similarities, correlations, or distances to prune a network based on the proximity between a pair of nodes. One of the most known, the Pathfinder algorithm [11], is used frequently because of its various mathematical properties, including the conservation of the triangle inequalities among a path of any number of links, the capability of modeling asymmetrical relationships, the representation of the most *salient* relationships present in the data, and the fact that hierarchical constraints in most cluster analysis techniques do not apply to Pathfinder Networks (PFNETs) [11].

A. Quirin (✉) and O. Cordón
European Centre for Soft Computing, Edf. Científico Tecnológico, Mieres, Spain
e-mail: [arnaud.quirin;oscar.cordon]@softcomputing.es

A. Abraham et al. (eds.), *Computational Social Network Analysis*, Computer Communications and Networks, DOI 10.1007/978-1-84882-229-0_9,

The resulting network could then be graphically represented using a network drawing algorithm. This methodology ensures that the network is represented in an aesthetic way, usually by adding several spatial constraints, such as the minimization of the edges crossings, the optimal distribution of the nodes over the space, and the minimization of the length of the edges. The final goal of this methodology is to generate a network having some properties. If the network represents the complex tissue of relationships between individuals or organizations, and if its goal is the comprehension of these relationships by a domain expert, we may be interested by some specific properties. For instance, the backbone could be better highlighted if it is drawn in the center, and minor links could be represented in the border of the map. The Kamada–Kawai [15] or the Fruchterman–Reingold [13] algorithm are usually applied for this task.

There are some kinds of Social Network Analysis (SNA) applications that could benefit from such methodology, giving to the domain expert or even a simple user a simple access to the information contained in these networks. One of these applications is co-citation network analysis. Co-citation network models depict the complex tissue of relationships occurring in the scientific literature. The graphical representation of these kinds of networks while preserving their information is still a challenge. However, some work has been done using *scientograms* [8, 9], visual representations showing the spatial distribution of the scientific actors in a given domain, wherein these actors can be as diverse as scientific categories, authors, journals, or papers. They show also additional information about the relationships between them, for instance, the proximity between two scientific authors. Because of the complexity of the domain they aim to represent and, more specifically, the fact that virtually all the scientific actors are connected together, these maps usually contain a large number of links and are really dense and hard to be directly represented.

The said methodology has already been described in the literature in the case of the scientograms [27]. But this methodology suffers from some drawbacks. First, the Pathfinder algorithm is very slow, avoiding the representation of the maps in an online way. Secondly, even if the Kamada–Kawai algorithm is the most used network drawing algorithm in this topic [19], it suffers from some aesthetic problems. In fact, this algorithm does not have an explicit procedure, either to avoid edge crossings, or to fill properly the full space allocated for the drawing. In the literature, the slowness of the Pathfinder algorithm has already been solved using a fast variant [24], but no algorithm has yet been proposed to overcome the drawbacks of the visualization of the Kamada–Kawai algorithm.

In fact, for the analysis of scientograms, the drawbacks of the Kamada–Kawai algorithm could prevent an expert from an optimal interpretation of the relationships taking place inside the considered scientific domain. For instance, edge crossings can make a node, and its labels overlap, thus avoiding a good reading of the map. Another point is the absence, in the Kamada–Kawai algorithm, of specific spatial constraints to avoid the links going back to the center of the map. Spatial artifacts, such as nodes appearing close together even if they are spatially separated by several links, could convey false or misinterpreting information.

In this chapter, we propose a new drawing algorithm to overcome these drawbacks. The structure of the current contribution is as follows. In Section 2, we review the existing methodology to design scientograms. In Section 3, we describe our proposal. In Section 4, some experiments will be shown. Finally, some concluding remarks are pointed out in Section 5.

9.2 A Methodology to Generate Scientograms

The achievement of a vast scientogram is a recurrent idea in the modern age. In 1998, Chen [8, 9] was the first researcher to bring forth the use of PFNETs in citation analysis. This is due to the fact that scientograms are the most appropriate means to represent the spatial distribution of research areas, while also affording information on their interactions [26]. Taking the latter as a base, Moya-Anegón et al. [21] proposed a method for the visualization and analysis of vast scientific domains using the ISI[1]-JCR category co-citation information. They represented it as a social network, simplified that network by means of the Pathfinder algorithm considering $q = n-1$ and $r = \infty$, and graphically depicted its layout using the Kamada–Kawai algorithm [15], thus getting a structural model of the scientific research in a vast domain. Note that $r = \infty$ and $q = n - 1$ are the common parameter values when Pathfinder is used for large domains scientogram generation. These values are very advantageous for large network pruning [10].

The different method stages are briefly described as follows. The last step is the one replaced by our proposal.

9.2.1 Category Co-citation Measure

Co-citation is a widely used and generally accepted technique for obtaining relational information about documents belonging to a domain. Because we strive to represent and analyze the structure of vast domains, whether they be thematic, geographic, or institutional, we fall back on ISI-JCR co-citation categories [21] as a tool.

Hence, once the rough information of the ISI-JCR co-citation for the categories present in the domain to be analyzed is obtained, a co-citation measure CM is computed for each pair of categories i and j as follows:

$$CM(ij) = Cc(ij) + \frac{Cc(ij)}{\sqrt{c(i) \cdot c(j)}} \tag{9.1}$$

where Cc is the co-citation frequency and c is the citation frequency.

[1] Currently registered as *Thomson Scientific*.

Notice that the aim of this scientogram generation method is that the final scientogram obtained is a tree. Hence, in order to avoid the existence of cycles in the pruned network, the considered measure of association adds the normalized co-citation (divided by the square root of the product of the frequencies of the co-cited documents' citations [25]) to the rough category co-citation frequency. In this way, the network weights become real numbers, allowing us to create small differences between similar values for the co-citation frequency, thus avoiding the occurrence of cycles and achieving the optimal prune of each link considering the citing conditions of each category.

9.2.2 Network Pruning by Pathfinder

Then, the Pathfinder algorithm is applied to the co-citation matrix to prune the network. We should take into account the fact that the networks resulting from citation, co-citation, or term co-occurrence analysis are usually very dense, when the categories are used as the unit for each node. Because of this fact, and especially in the case of vast scientific domains with a high number of entities (categories in our case) in the network, Pathfinder is parameterized to $r = \infty$ and $q = n - 1$, in order to obtain a schematic representation of the most outstanding existing information by means of a network showing just the most salient links. In general, the weights of the links of the co-citation matrix belong to $\mathbb{R}$ and are all different, so the final result of the Pathfinder algorithm is a tree. To perform this step, the MST-Pathfinder algorithm, a quick version of the original Pathfinder algorithm based on *Minimum Spanning Trees*, is used [24].

9.2.3 Network Layout by Kamada–Kawai

Kamada–Kawai algorithm [15] is then used to automatically produce representations of the pruned network resulting from the Pathfinder run on a plane, starting from a circular position of the nodes. It generates social networks with aesthetic criteria such as common edge lengths, forced separation of nodes, building of balanced maps, etc. Nevertheless, some criteria are not directly satisfied in the Kamada–Kawai algorithm. This is the case of a number of crossed links: in fact many links crossings appear making a lot of nodes overlap, and making the reading of the map harder. Another point is the fact that edges can go backwards to the center of the map, putting close two nodes linked by a long path. This can give a false impression of closeness to the expert because of the spatial distribution of the nodes. These are two of the main drawbacks we observed while using the Kamada–Kawai visualization applied to co-citation networks.

An example of the render of the Kamada–Kawai applied on the co-citation network of Europe is shown in Fig. 9.1.

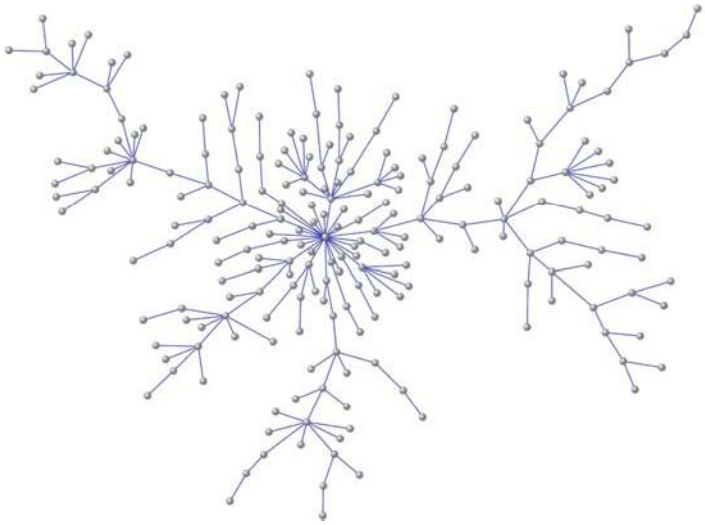


Fig. 9.1 An example of a scientogram corresponding to the Europe scientific domain in 2002

9.2.4 Advantages and Drawbacks of Our Methodology

Since its proposal in 2004, this methodology has been applied to compare the structure of vast scientific domains [20, 27], to the macro and microstructural analysis of a specific domain [18, 19], and even to study their evolution through time [28, 29].

Hence, it is actually a very powerful tool due to its summarization capability as well as its simplicity to represent the relational information linked through a series of intelligible sentences that make easier the comprehension, analysis, and interpretation of a scientific domain. However, the critics that can be made on the Kamada–Kawai algorithm about the crossed links make it difficult to apply for the generation of scientograms in an automatic way, as a human post-processing is currently needed to avoid the crossings and move the labels in order so that they can be read clearly. We aim to solve these drawbacks using a new visualization algorithm, as we will see in the remainder of this chapter.

9.2.5 The Use of Other Network Layout Algorithms for Social Networks Drawing

Apart from the Kamada–Kawai algorithm, an extensive number of graph drawing algorithms have been already published in the literature [2, 3]. But very few of them have been devoted to the drawing of social networks because of the high complexity of this kind of networks [19].

In [6], the authors have chosen a radial-based network drawing algorithm to represent a collaboration network, and they named it *Visone*. In this kind of radial drawing, the nodes are placed in concentric circles, thus highlighting the role of the central elements and pushing away the less important elements on the boundaries. Other studies on the same kind of drawing algorithms for citation networks or co-authorship networks are presented in [5, 7]. Although the obtained representation is nice, the placement of the nodes are too much constrained, thus the length of the links can suffer from meaningless growing, making the map harder to understand.

In general, force-directed layout algorithms are studied in the literature to deal with social networks. The study of Katz and Stafford [16] present the results obtained with the Fruchterman–Reingold algorithm compared to the Kamada–Kawai algorithm to visualize the American Federal Judiciary network. The Fruchterman–Reingold algorithm is also used in [12] for the visualization of the evolution of social networks over time. In general, this algorithm gives interesting results as it reveals clusters of nodes. For our application to the co-citation networks, both force-directed layout algorithms were tested, but the Kamada–Kawai algorithm was preferred due to its aesthetic behavior [19].

In conclusion, a few different drawing algorithms are studied in the literature, while the need of a comprehensible and aesthetic drawing algorithm is growing as social networks become more and more complex.

9.3 Overview of the Algorithm

This section describes our drawing algorithm. As said, our methodology ensures that the result of the Pathfinder algorithm, the network we have to draw, is a tree. Thus, to develop our algorithm, we took as a base the tree visualization algorithm presented in [22], and extend it to make it applicable on scientograms. The basic version of the algorithm is first presented, then several variants are discussed for the specific case of scientogram design.

9.3.1 Main Algorithm

To ease the understanding of our proposal, some preliminary terminology is first introduced. In the following, we consider an ordered tree T, in which each node N has a parent P = PARENT(N), except the root node R = ROOT(T). CHILDREN(N) is the set of nodes having node N as their parent. ASCENDANT(N) is the chain of nodes from node N to the root node R, defined as { N, PARENT(N), PARENT(PARENT(N)), ..., R }. SUBTREE(N) is the subtree having node N as its root. SIZE(N) is equal to the number of nodes in SUBTREE(N), including its own root. For instance, SIZE(N) is equal to 1 for a node having no children; 2 for a node having one child; etc. LEVEL(N) is the number of nodes in the set

ASCENDANT(N). For instance, LEVEL(N) is equal to 1 for the root node; 2 for any of the children of the root node; etc. DEPTH(N) is the maximum value for LEVEL(M) for any node M in the tree SUBTREE(N). For instance, DEPTH(N) is equal to 1 for a node having no child; 2 for a node having any number of children, with none of them having children; etc. By convention, we will also use the notation ROOT(T) to define the node having the lowest level in a subtree T.

The algorithm is named *Vmap-Layout* because of the kinds of maps it draws, i.e., *Visual Science Maps*, another name for scientograms. It is divided itself into three sub-functions, which are called in a sequential way. The first sub-function, *Attribute Computation*, computes for each node the attributes needed for the remainder of the algorithm. The second subfunction, *Node Positioning*, is a recursive function aiming to compute the coordinates of each node. The third one, *Node Relocation*, adjusts the location of the nodes according to some specific criteria, thus improving the final visualization. The different sub-functions are detailed in the following sections.

9.3.2 Attribute Computation

Within the first sub-function, we compute several attributes assigned to each node: SIZE(N), LEVEL(N), and DEPTH(N). These attributes will be used later to facilitate the generation of the coordinates of each node and to improve the runtime of the algorithm. The first step is to select a root node, the one that will be printed in the center of the map. It will be used to compute some specific attributes that cannot be computed without the definition of a root node. The network generated by the Pathfinder algorithm does not selfcontain any root node; hence we have to use an additional technique to select one. There are many ways to select a center in a graph. Many of them are described by Bavelas [4] and Parlebas [23]. The one used here, that gives good visual results, is the *deliverer criterion*: we compute the sum of the distances between any node and all the others, and we take as the root the one having the smaller value. Once we have selected the root node R, a number of other attributes can be computed. Giving R, we can assign to each node N the values corresponding to SIZE(N), LEVEL(N), and DEPTH(N). As the tree is represented in memory using lists of children, the time complexity of all these operations is $O(n)$ where $n = SIZE(R)$. The complexity of *Attribute Computation* is thus $O(n)$. The sub-function is outlined in Fig. 9.2.

1. Compute the root node using the *deliverer criterion*: compute the sum of the distances between a node and all the others, and take as a root the one having the smaller value.
2. Assign to each node N the values given by SIZE(N), LEVEL(N) and DEPTH(N).

Fig. 9.2 The *Attribute Computation* sub-function

9.3.3 Node Positioning

Using the second sub-function, the algorithm fixes the location of each node as a pair of 2D coordinates. To do so, the global idea is to fill as much space as possible. The tree is drawn from the root node to the leaves, and the algorithm runs in a recursive way: the root node is drawn in the center of the map and at each iteration the algorithm draws all the nodes N having the same level $\mathrm{L} = \mathrm{LEVEL}(N)$. The algorithm starts by selecting a region of the empty space in which it can draw the tree (we will call this region *initial polygon* in the following). Then, it assigns the root node to the center of this polygon, it divides the initial polygon into several slices (as many as the number of children in the root node), and it assigns one child of the root node to each slice. After the application of this function, the coordinates of the center of the polygons are assigned to each node.

There are several ways to design the initial polygon. As the whole tree will lie inside the initial polygon, the later will determine the final shape of the full map. This shape could be a square, a circle, or any other n-sided polygon. Figure 9.6 shows some possibilities for the initial polygon. The assignation of a center C to a given polygon could also be done in several ways. Some techniques are shown in Fig. 9.8. Lastly, the division of a given polygon into different slices can also been done in several ways, which are detailed in the next sections.

Once the coordinates of the first level of the tree are fixed, the algorithm starts again, considering each of the slices as a new polygon and the corresponding node N as the root of a new subtree $S = \mathrm{SUBTREE}(N)$. Once no child has been found, the algorithm stops. The sub-function is outlined in Fig. 9.4. As we are using a recursive function based on the children on a given node, the complexity of *Node Positioning* is thus $O(n)$.

At the end of this sub-function, all the nodes of the tree have been assigned to a pair of coordinates in the 2D-space. An example of the execution of this sub-function on a tree is shown in Fig. 9.3.

9.3.4 Node Relocation

The goal of the third sub-function is only aesthetic. At the end of the application of the previous sub-function, some graphical elements, such as the nodes or the text labels, can overlap. By relocating the nodes, we improve the placement of the overlapping elements. Here, having the coordinates of the nodes generated previously, the algorithm fixes the final location of each node to avoid these overlaps as much as possible. We say that two nodes overlap when they are too close to each other, according to a distance defined by the expert, and we call them *problematic* nodes. Problematic nodes cause text labels not to be read in a clear way, and reduces in general the readability of the map. The main idea of this function is to apply a node relocation process in which the problematic nodes are moved according to a repulsive force depending on the surrounding nodes, like if they were connected with

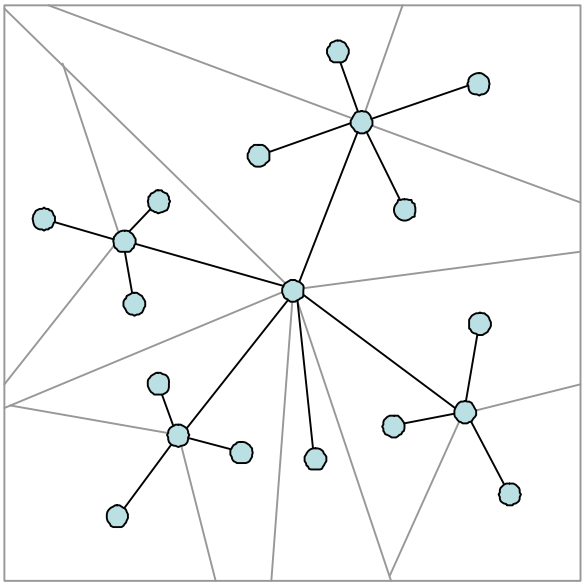


Fig. 9.3 An example of the execution of the *Node Positioning* sub-function

1. Let P, a 2D-space region in which to draw the tree, and T, a tree.
2. Choose a central point C in P and assign to this point the root of the tree R = ROOT(T).
3. If CHILDREN(R) is empty, stop.
4. Divide P in different slices (*sub-polygons*), giving as many sub-polygons that the number of children of R. Let R_i be a child of R, the area of the corresponding sub-polygon P_i should be proportional to SIZE(R_i).
5. For each child R_i of R, run *Node Positioning* on the region P_i and the tree R_i.

Fig. 9.4 The *Node Positioning* sub-function

repulsive springs. To avoid deadlocks in some cases (e.g., when a node is located exactly between other two nodes and at an equal distance), the problematic nodes are also slightly moved in a random direction, until they met a criterion set by the expert.

The sub-function is outlined in Fig. 9.5. As the time complexity of the KD-Tree preprocessing is $O(n \cdot \log(n))$ and the time complexity of the KD-Tree search for one node is $O(\log(n))$ [1], the time complexity of *Node Relocation* is $O(n \cdot \log(n))$. Thus, the complexity of the full algorithm is $O(n \cdot \log(n))$.

At the end of the third function, the final coordinates of the nodes have been computed. During the final drawing of the nodes, additional improvements could be made in order to improve the aesthetic aspect of the map. For instance, nodes and label sizes can be varied depending on their depth in the tree to better highlight the center of the map.

1. Apply a KD-Tree technique to compute the distance between all the nodes of T.
2. Select only the problematic nodes, i.e., the nodes close enough according to a criterion defined by the expert.
3. For each node of this set, do:
   - Apply a repulsive strength δ and move the node along this force.
   - Move it around its final position using a small random distance in the interval $[-\sigma,\sigma]$.
4. Execute again the *Node Relocation* sub-function until a given amount of iterations defined by the expert has been reached.

Fig. 9.5 The *Node Relocation* sub-function

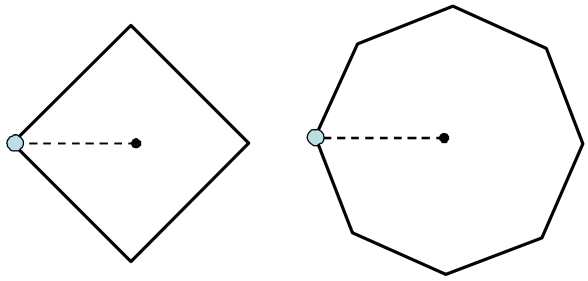


Fig. 9.6 Initial polygons with four or eight sides

9.3.5 Selecting Different Initial Polygons

During the initialization of the Vmap-Layout algorithm, we have to select the initial polygon, within which the full tree has to be drawn. This polygon encloses all the layout and its shape will determine the global shape of the drawing.

In order to improve the general aspect of the final map and to customize the result for several uses (paper or online drawing), several options can be used. The definition of this shape is controlled by an expert parameter, giving the number of sides the initial polygon should have (see Fig. 9.6). The larger this value, the more circular the shape will be, but the slower will the algorithm run. This is due to the fact that, as at later stages, the computation of the areas and the angles of a sub-polygon would be more complex. This number of sides does not change in any manner the further execution of the algorithm but has only an aesthetic aspect.

Once the initial shape is designed, it is used to draw the initial tree, composed of its root and of all its first-level children (see Fig. 9.7). Any shape surrounding the graph could be used. For the application to co-citation networks, we opted by a circle because the SCImago Research Group[2] experts, with whom we collaborate, prefer this shape. Thus, in order to have a good compromise between time and aesthetic,

[2] http://www.scimago.es/

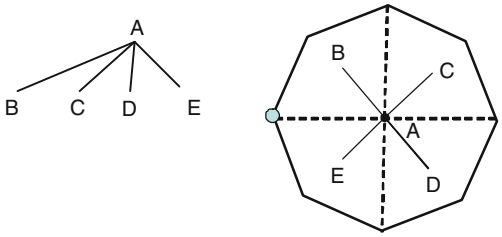


Fig. 9.7 On the *left*, the tree to draw. On the *right*, the initial polygon and the *center* in black. The polygon is first divided into four slices, because the node A has four children, and then we assign the corresponding nodes to the corresponding sub-polygons. The little circle is the starting point giving the direction of the polygon assignment

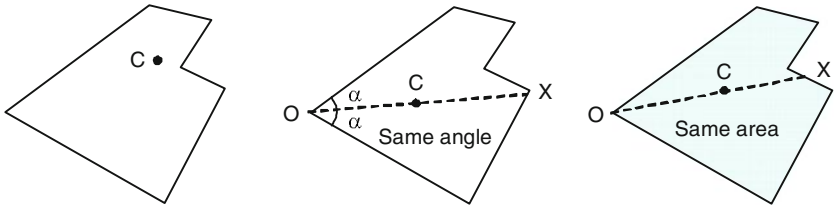


Fig. 9.8 Centers positioned with the *Center of Mass*, the *Angle-Based Central Point* and the *Area-Based Central Point* methods

a value of 15 sides seems to be well suited. Larger values than 30 will unnecessarily increase the run-time and smaller values than 12 would give an impression of discontinuity.

9.3.6 Selecting Different Ways to Compute the Central Point of a Polygon

For each polygon, a central point has to be selected to become the starting point of the next sub-tree to print (see step 2 in Fig. 9.4). We have explored at least three different methods to choose the central point C of a polygon P (see Fig. 9.8). The first method, called "*Center of Mass*," takes for C the center of gravity of P. This center is defined in any case, but suffers from two problems: it could be outside of the polygon (this can occur when the polygon is nonconvex) and a polygon with a lot of segments could attract the center far away from the *natural* center of the polygon. The second method, called "*Angle-Based Central Point*," uses the angle to compute the central point. For a given polygon P, we first select a point on its border, that we call origin O. This origin O has itself to be defined by some methods. For instance, the origin could be the center of the parent polygon (the one used to generate the current polygon), or the left-most point of the polygon. We then draw a line dividing the angle O in two equal parts. Then we take the middle point of this line as

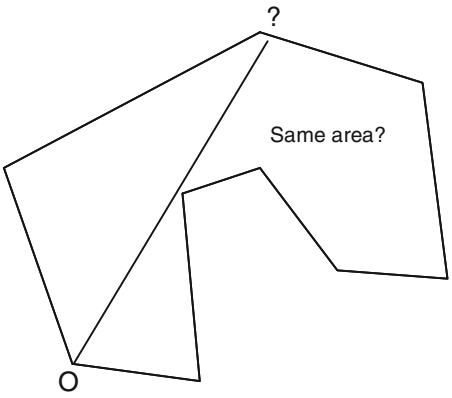


Fig. 9.9 A problem that can occur with the *Area-Based Central Point* method: the two areas cannot be made equal

the center C of the polygon. The third method, called "*Area-Based Central Point*," applies the same procedure, but by dividing the polygon into two parts having the same area.

Our tests have shown that the *Area-Based Central Point* is the best method. However, some problems could occur for some specific shapes of polygons in which it is impossible to divide a sub-polygon into two areas of equal sizes, pushing the central point C outside the polygon. This can happen when the polygon has internal angles greater than pi (see Fig. 9.9), when the chosen initial polygon is nonconvex or when the structure of the tree is quite uncommon. That is why other methods are provided.

For the *Angle-Based Central Point* and the *Area-Based Central Point* methods, once the line dividing the polygon in two parts has been determined, we still have to place the point C over this line. The usual way is to use the middle of the line, as described in the first paragraph of this section. But other values for the measure of the distance OC have been explored. This distance has a direct influence on the length of the edges and the location of the next sub-polygons, and playing with this value can lead to interesting results on the final drawing. An expert parameter has been set for this measure and is named *Cutpoint Value*. It is the ratio between the distance OC and the distance OX (see Fig. 9.8). With a small value for this parameter, we get maps where the centers are close among them, and with a larger value, we get maps where the centers are more far away among them. Several values for this parameter have been tried, such as 0.25, 0.5, and 0.6. The best results have been obtained with values lower or equal to 0.5.

9.3.7 Selecting Different Dividing Slice Methods

The way to divide a polygon into different slices (see step 4 in Fig. 9.4) will determine the respective areas for the drawing of the next sub-polygons. Small areas should be allocated to small sub-trees, whereas larger areas should be allocated to larger sub-trees. This operation can be achieved by at least two methods (see

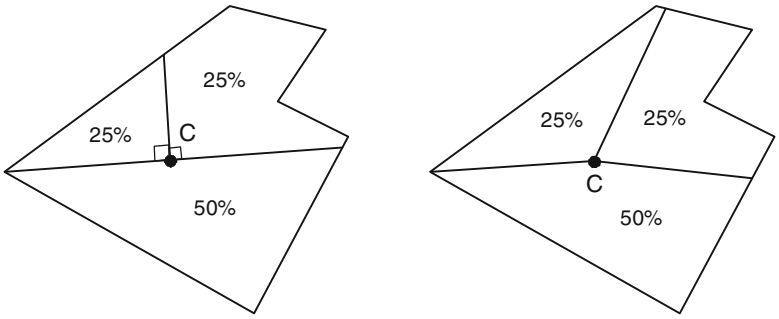


Fig. 9.10 The result of the dividing using the *Angle-Based Dividing* (*on the left part*) and the *Area-Based Dividing* (*on the right part*) methods

Fig. 9.10). With the first method, called *Angle-Based Dividing*, the algorithm defines the size of the slices in order the angle around the center C is proportional to the size of each sub-tree R_i. With the second method, called *Area-Based Dividing*, the algorithm defines the size of the slices in order the area around the center C is proportional to the size of each sub-tree R_i. In many scientograms generated using real world data, the second method does not work properly because the non-convex shapes of the polygons make the finding of a percentile using the area an impossible problem (see Fig. 9.9). Therefore, in our application, we always used the *Angle-Based Dividing* method.

Any use of the two previously described methods needs a measure defined for each node in order to compute the proportion in percentage allocated to the corresponding sub-polygon. The simplest way, called the *Sub-Size-Based Ratio Computation*, is to take the size of each node (defined by the SIZE(T) attribute), i.e., the number of nodes in a subtree T, to compute a proportional ratio assigned to the children of T. Then, this ratio is used as a percentage to compute the size of all the slices of the corresponding sub-polygons. Nevertheless, this method suffers from a lack of customization possibilities by the expert.

Another method called the *Sub-Depth-Based Ratio Computation* has been explored. It uses the depth of the trees, defined by the DEPTH(T) attribute, to modify the proportion allocated to each slice depending on whether they are close or far away from the center of the map. The expert has to set two additional parameters, the proportion given for the allocation of the slice of the lowest level, corresponding to the initial root of the tree (this value has been named ”START”), and the proportion given for the allocation of the slice of the deepest level, corresponding to a given leaf of the tree (this value has been named ”END”). Because only these two values, START and END, have to be specified by the expert, the remaining values used to fix the proportion of the intermediate levels are computed using a linear regression. Using another point of view, the START value fixes the behavior of the nodes close to the center of the map (or the backbone), which are the most important ones, and the END value fixes the behavior of the minor nodes shown

in the periphery. These parameters are thus useful for our application of co-citation networks. We have obtained the best results using 0.5 and 0.25 for the respective values of START and END.

The behavior of the *Sub-Depth-Based Ratio Computation* method is very simple. It allows us to constrain the angle of the links to go only forward when we are close to the center of the map (see Fig. 9.11). In fact, when drawing scientograms, having edges going mainly forward gives a better representation as nodes located far away in terms of number of edges are spatially dissociated. That is why we selected the *Sub-Depth-Based Ratio Computation* method as the default one for our experimentations.

To show how the proportion of the slices are computed using the *Sub-Size-Based Ratio Computation* and the *Sub-Depth-Based Ratio Computation* methods, an example is presented in Fig. 9.12 using a small tree. For some selected nodes, the values are computed for both methods. The proportion is always computed as a

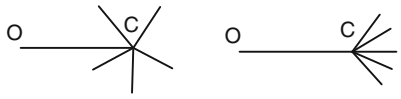
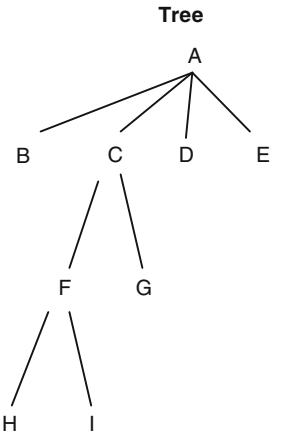


Fig. 9.11 On the left, the normal behavior in which all the space is used to compute the size of each slice. On the right, a modified behavior in which a constraint is applied before computing the size of each slice, allowing us to direct the network in a given way



(1) Normal behavior

Measure : the 'sub_size' attribute

Node	Measure	Proportion
B	1/8	12.5%
C	5/8	62.5%
F	3/4	75%
H	1/2	50%

(2) Modified behavior

Measure : the 'sub_depth' attribute

Node	Measure	Proportion
B	2/2	the value 'START'
C	2/4	0.5 'START' + 0.5 'END'
F	3/4	0.25 'START' + 0.75 'END'
H	4/4	the value 'END'

Fig. 9.12 Different ratio computation methods for dividing the slices: an example of a tree (*left*); the proportions obtained using the *Sub-size-Based Ratio Computation* method (*top*); and the proportions obtained using the *Sub-Depth-Based Ratio Computation* method (*bottom*)

ratio between two numbers, the current attribute of the node (respectively the SIZE and the DEPTH) and the maximum value for this attribute (so, the maximum size of the current sub-tree or the maximum depth if the second measure is considered). Note that the ratio computation method is totally independent from the method of dividing these slices, i.e., the ratio can be used independently with the *Area-Based Dividing* or the *Angle-Based Dividing* methods.

9.3.8 Details on the Node Relocation *Function*

Once the coordinates of the nodes have been found by the *Node Positioning* sub-function (see Fig. 9.4), a node relocation stage occurs in order to improve the location of the overlapping nodes. This stage is done by the *Node Relocation* sub-function (see Fig. 9.5).

The goal of this sub-function is to identify the nodes that are too close, according to an expert criterion, and to move them randomly in order to avoid the overlapping of the nodes. The process is iterated several times until a perfect configuration is found by the algorithm. Because of the cost of the computation of the distance between nodes, and since this process has to be iterated, we use a KD-Tree technique to compute these distances. A KD-Tree allows the computation of the distance between a set of nodes in order that we can quickly know which set of nodes is closer to any given node.[3]

The *Node Relocation* sub-function works as follows. A parameter named *radius* is defined by the expert to specify the minimum allowed distance between two nodes. Only the coordinates of any node having a smaller or equal distance to this radius will be modified by the algorithm, while the coordinates of the nodes located at a greater distance will not be changed. Two additional parameters are defined, the *spring-strength* δ giving the strength of the movements during the relocation of the nodes and the *random-strength* σ giving the quantity of randomness applied to the nodes that have to be relocated.

From each node N of the set defined by the *radius* parameter, we apply a force defined as the sum of all the repulsive forces generated by the nodes close to N, multiplied by the value defined by the *spring-strength* δ parameter (a repulsive strength), and add a random value chosen into the interval $[-\sigma,\sigma]$ to it. Figure 9.13 shows how the repulsive forces generated by all the surrounding nodes are applied on a given node, and how this node is moved. This is done until a given number of iterations have been completed. A higher value for this parameter can be used to establish the convergence of the coordinates of the nodes, but at the cost of slowing down the process. We have obtained good and fast results with a value of 100 iterations.

[3] The library we have used for this process is called the ANN Library [17]. The advantage of this library is that an approximate distance computation is used in order to speed up the process, provided the fact that knowing the exact values of the distances is not important, which is the case in our application.

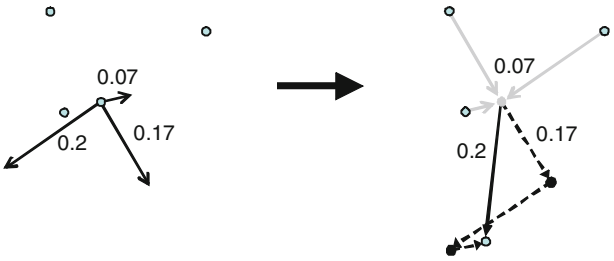


Fig. 9.13 An example of the modification of the coordinates of a node after applying the *Node Relocation* function

The *spring-strength* δ parameter is used to define the step size of movement applied to the node. A small value moves the nodes slowly through the iterations of the algorithm, while a bigger value allows sudden changes of the coordinates of the nodes. A value of 0.10 was used in our experiments. The *random-strength* defines the quantity of randomness applied to the location of the node at the end of each iteration. A value of 0 disables any randomness during the movement of the nodes. A value of 0.05 was used.

9.4 Experiments

In this section, we will show the results obtained on some scientograms. The data are directly extracted from a database of co-citation measures for Europe, generated in 2002. Some specific criteria are set in order to select a subset of the whole database restricted to four world regions: Europe, the USA, Spain, and Cuba. The resulting file encodes a fully connected network, with labeled nodes and weighted links, ready to be pruned. Thus, the first step is to use the MST-Pathfinder algorithm to prune these networks in order to get trees. The computing time for this step is roughly 9 ms on an Intel dual-core Pentium 3.2 GHz with 2 GB of memory.

The next step is to print the maps using the Vmap-Layout algorithm. The algorithm has been written in C++, and compiled on Linux with the GNU `GCC` compiler with the `-O3` option. The computing time using the Vmap-Layout algorithm is roughly 71 ms. The main parameters used are as follows. The initial polygon has 15 sides. The central point has been computed using the Angle-Based Central Point method. The Cutpoint Value has been set to 0.5. The slices have been divided using the Angle-based dividing method. Finally, the Sub-Depth-Based Ratio Computation has been used to position the central point in the polygon.

A comparison has been performed using the Kamada–Kawai algorithm. For this purpose, we have used the GraphViz library. GraphViz is an open source network drawing software, freely provided by AT&T Labs, and available at: http://www.graphviz.org/. It integrates the Kamada–Kawai algorithm in the form of the *neato* utility. This utility exports in diverse graphical formats a description of a network

done with a proprietary language, the DOT language [14]. Thus, the second step was to convert the previous network in the format accepted by this library. The computing time using the Kamada–Kawai algorithm is 0.6 ms.

The last step is to generate the graphical output from the DOT description using *neato*. Actually, as these maps are designed for on-line consultation, the Scalable Vector Graphics (SVG) format was chosen. The time to generate the SVG image is 1,300 ms. The following command was used to generate this map, once the library is installed:

```
dot -Kneato -Tsvg -o Europe.svg Europe.dot
```

The final results, for the Vmap-Layout and the Kamada–Kawai algorithms are shown in Figs. 9.14 and 9.15 for Europe; in Figs. 9.16 and 9.17 for the USA; in Figs. 9.18 and 9.19 for Spain; and in Figs. 9.20 and 9.21 for Cuba.

Fig. 9.14 The scientogram of Europe in 2002, drawn with the Kamada–Kawai algorithm

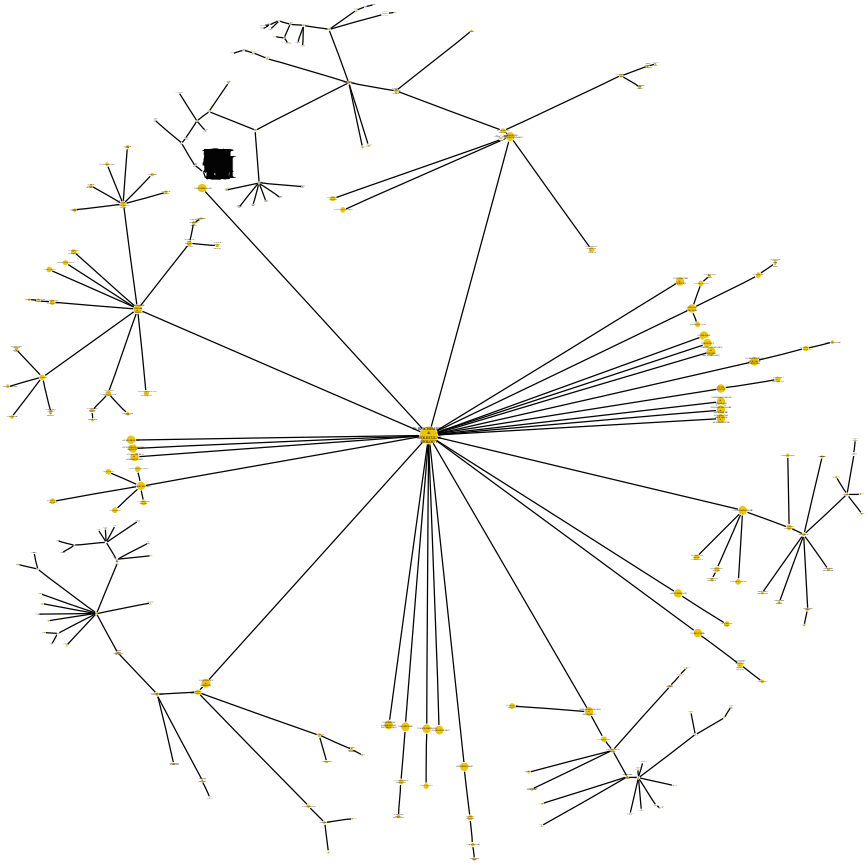


Fig. 9.15 The scientogram of Europe in 2002, drawn with the Vmap-Layout algorithm

Several remarks can be made from these pictures. First, we have to notice that the Vmap-Layout algorithm avoids the edge crossings as expected, as each sub-tree is drawn in its own sub-space. Secondly, the nodes connected to the central node are properly spaced, and aligned on its own circle, allowing an expert to read properly the labels. In the case of the Kamada–Kawai maps, the reading of the top-level labels is not so clear. Finally, all the space available in the figure is properly filled with the Vmap-Layout, giving more space to the larger sub-trees.

9.5 Conclusion

The Vmap-Layout algorithm is an effective and a fast technique for the representation of co-citation networks. Our algorithm can print real-world networks in an

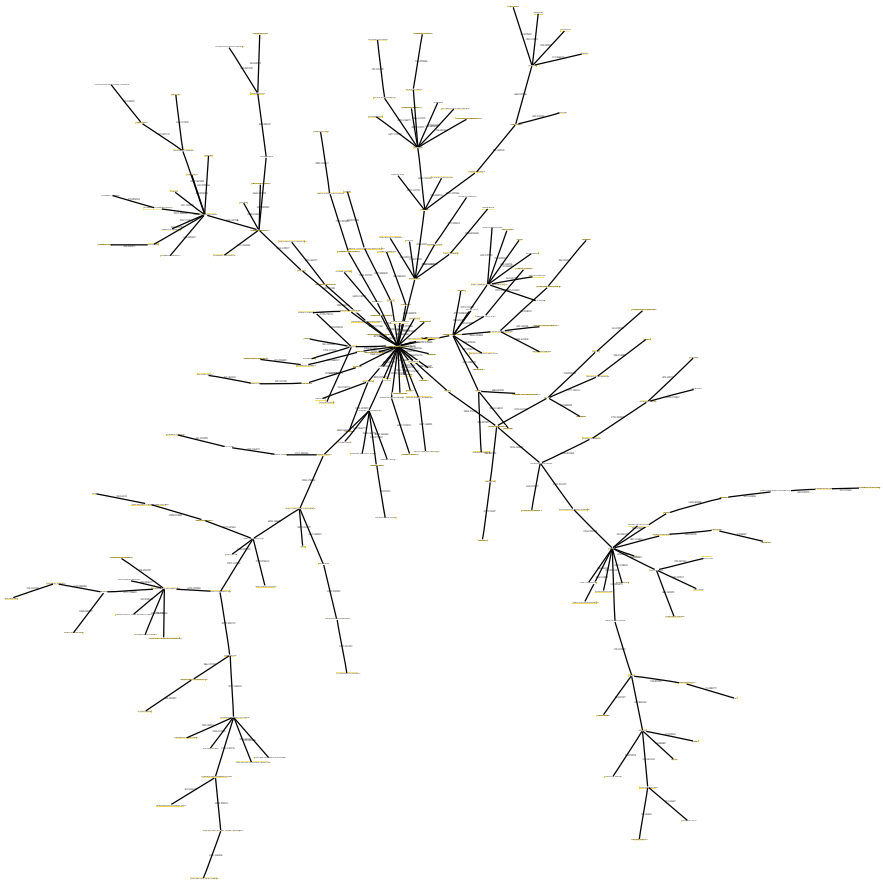


Fig. 9.16 The scientogram of the USA in 2002, drawn with the Kamada–Kawai algorithm

aesthetic way, highlighting the backbone and pushing the less important links to the boundaries. Several variants have been described, allowing an expert to tune the representation depending on his needs.

We are currently investigating other improvements of this algorithm. One option is to allow it to use extra space over the polygons in order to reduce the white space between the edges. Another option is the use of different techniques for the node relocation, for instance, based on the simulated annealing metaheuristic, to find a better positioning of the nodes. We are also planning some experiments on other kinds of real data sets, including larger networks.

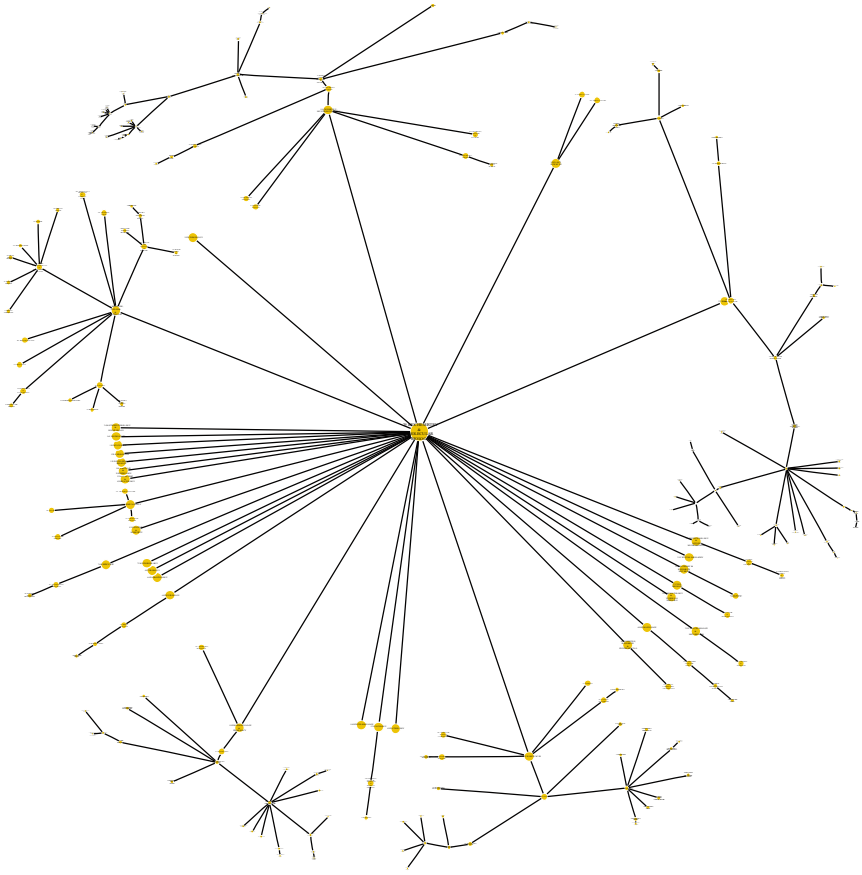


Fig. 9.17 The scientogram of the USA in 2002, drawn with the Vmap-Layout algorithm

Fig. 9.18 The scientogram of Spain in 2002, drawn with the Kamada–Kawai algorithm

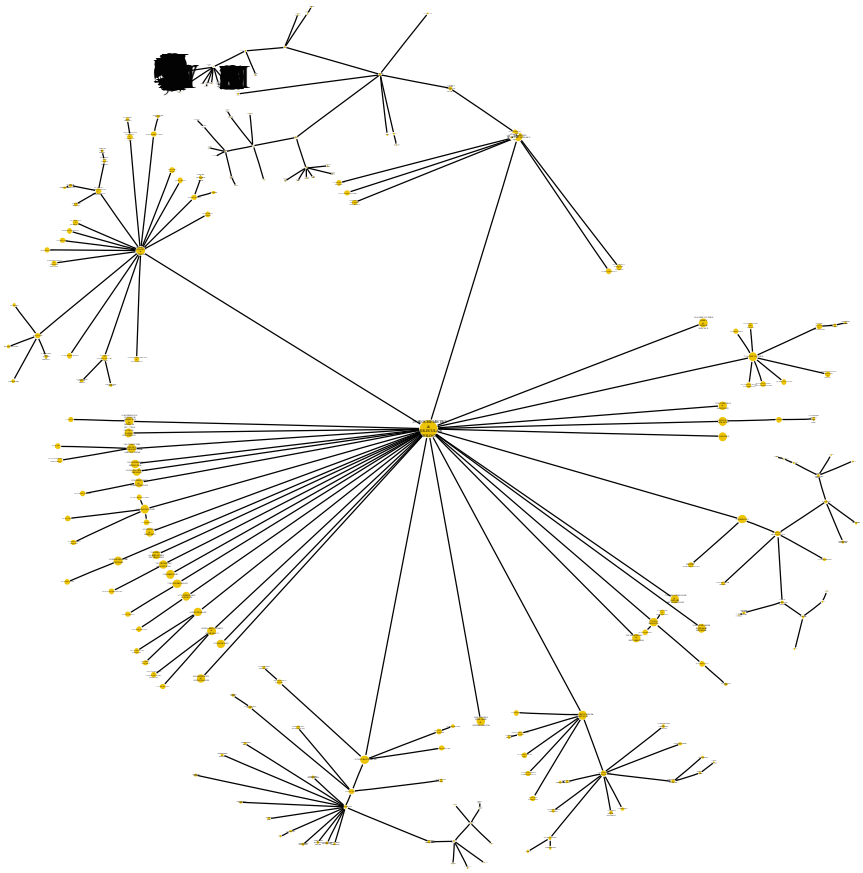


Fig. 9.19 The scientogram of Spain in 2002, drawn with the Vmap-Layout algorithm

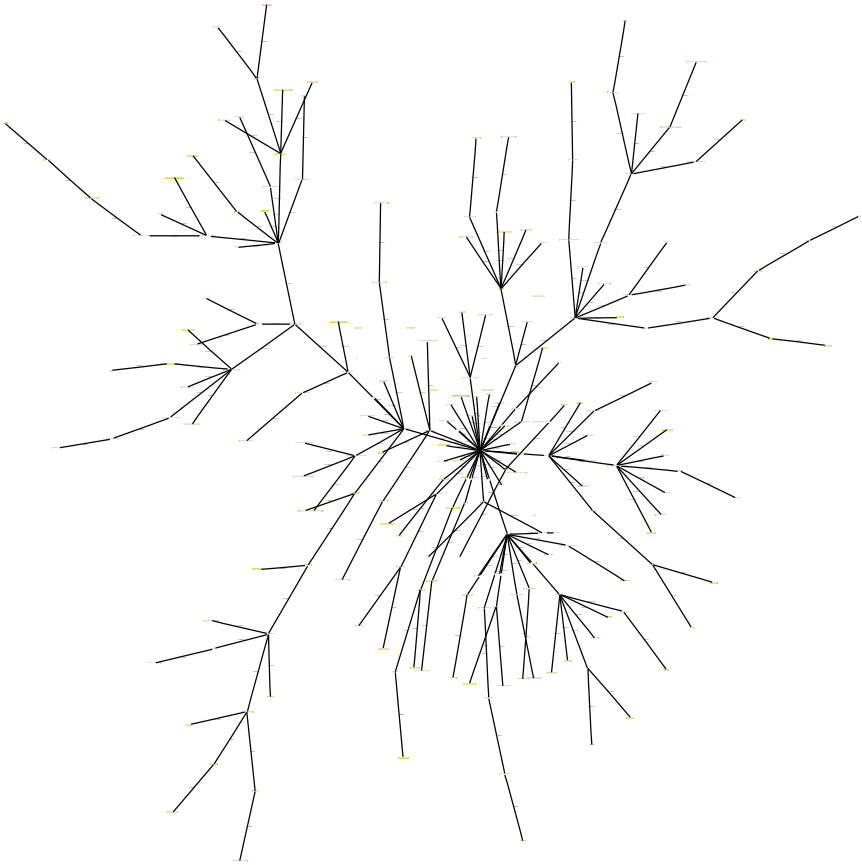


Fig. 9.20 The scientogram of Cuba in 2002, drawn with the Kamada–Kawai algorithm

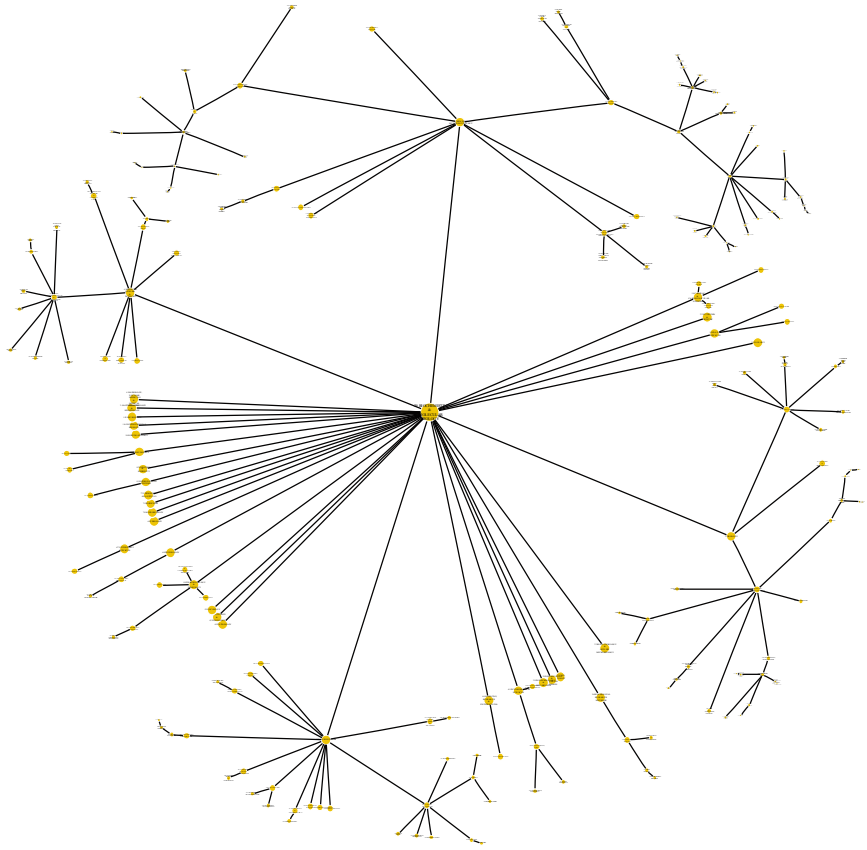


Fig. 9.21 The scientogram of Cuba in 2002, drawn with the Vmap-Layout algorithm

References

1. Arya S, Mount DM, Netanyahu NS, Silverman R, Wu AY (1998) An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J ACM 45(6):891–923
2. Battista GD, Eades P, Tamassia R, Tollis I (1994) Algorithms for drawing graphs: An annoted bibliography. Comp Geo Theor Appl 4(5):235–282
3. Battista GD, Eades P, Tamassia R, Tollis I (1999) Graph Drawing. Prentice Hall, Upper Saddle River, NJ
4. Bavelas A (1951) Réseaux de communications au sein de groupes placés dans des conditions expérimentales de travail, Les sciences de la politique aux États- Unis. Armand Colin, Paris
5. Brandes U, Erlebach T (2005) Network analysis: Methodological foundations, Vol. 3418 of Lecture Notes in Computer Science. Springer, Berlin
6. Brandes U, Kenis P, Wagner D (2003) Communicating centrality in policy network drawings. IEEE Trans Vis Comput Gr 9(2):241–253
7. Brandes U, Wagner D (2003) Graph drawing software, Visone – analysis and visualization of social networks. Berlin, pp 321–340
8. Chen C (1998) Bridging the gap: The use of pathfinder networks in visual navigation. J Vis Lang Comput 9:267–286
9. Chen C (1998) Generalised similarity analysis and pathfinder network scaling. Interact Comput 10:107–128
10. Chen C (2004) Information visualization: Beyond the horizon. Springer, Berlin, Germany
11. Dearholt D, Schvaneveldt R (1990) Properties of pathfinder networks. In: Pathfinder associative networks: Studies in knowledge organization. Ablex Publishing Corporation, Norwood, NJ, pp 1–30
12. Dynes SBC, Gloor PA, Laubacher R, Zhao Y, Dynes S (2004) Temporal visualization and analysis of social networks. In: North American Association for Computational Social and Organizational Science Conference (NAACSOS), Pittsburgh
13. Fruchterman TMJ, Reingold EM (1991) Graph drawing by force-directed placement. Software Pract Exper 21(11):1129–1164
14. Gansner ER, North SC (2000) An open graph visualization system and its applications to software engineering. Software Pract Exper 30(11):1203–1233
15. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. Info Process Lett 31(1):7–15
16. Katz DM, Stafford DK (2008) Hustle and flow: A social network analysis of the american federal judiciary. In: Annual Meeting of the The Law and Society Association, Montreal, Canada
17. Mount DM, Arya S (2006) ANN: A library for approximate nearest neighbor searching, version 1.1.1, online software available on http://www.cs.umd.edu/~mount/ANN/, released the 4/8/2006
18. Moya-Anegón F, Vargas-Quesada B, Chinchilla-Rodríguez Z, Corera-Álvarez E, González-Molina A, Muñoz-Fernández F, Herrero-Solana V (2006) Vis. y Anal. de la Estr. Sci. Esp.: ISI web of science 1990–2005 (in Spanish). El Profesional de la Información, 15(4):258–269
19. Moya-Anegón F, Vargas-Quesada B, Chinchilla-Rodríguez Z, Corera-Álvarez E, González-Molina A, Muñoz-Fernández F, Herrero-Solana V (2007) Visualizing the marrow of science. J Am Soc Info Sci Technol 58(14):2167–2179
20. Moya-Anegón F, Vargas-Quesada B, Chinchilla-Rodríguez Z, Corera-Álvarez E, Herrero-Solana V, Muñoz-Fernández F (2005) Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation. Info Process Manage 41:1520–1533
21. Moya-Anegón F, Vargas-Quesada B, Herrero-Solana V, Chinchilla-Rodríguez Z, Corera-Álvarez E, Muñoz-Fernández F (2004) A new technique for building maps of large scientific domains based on the cocitation of classes and categories. Scientometrics 61(1):129–145
22. Nguyen QV, Huang ML (2002) A space-optimized tree visualization. In: IEEE symposium on information visualization (InfoVis 2002), pp 85–92
23. Parlebas P (1972) Centralité et compacité d'un graphe. Math. et Sci. Hum. 39:5–26

24. Quirin A, Cordón O, Guerrero-Bote VP, Vargas-Quesada B, Moya-Anegón F (2008) A quick MST-based algorithm to obtain pathfinder networks. J Am Soc Info Sci Technol 59(12):1912–1924
25. Salton G, Bergmark D (1979) A citation study of computer science literature. IEEE Trans Prof Commun 22:146–158
26. Small H, Garfield E (1985) The geography of science: Disciplinary and national mappings. J Info Sci 11:147–159
27. Vargas-Quesada B, Moya-Anegón F (2007) Visualizing the structure of science. Springer, New York
28. Vargas-Quesada B, Moya-Anegón F, Chinchilla-Rodríguez Z, Corera-Álvarez E, González-Molina A, Muñoz-Fernández FJ, Herrero-Solana V (2008) Evolución de la estructura científica española: ISI web of science 1990–2005 (in Spanish). El Profesional de la Información 17(1):22–37
29. Vargas-Quesada B, Moya-Anegón F, Chinchilla-Rodríguez Z, González-Molina A (2007) Showing the essential science structure of a scientific domain and its evolution. Info Vis (in press)