Focus-based filtering + clustering technique for power-law networks with small world phenomenon

Mountaz Hascoët, François Boutin, Jérôme Thievre

Focus-based filtering + clustering technique for power-law networks with small world phenomenon

François Boutin[a], Jérôme Thièvre[b] and Mountaz Hascoët[c]

[a] LIRMM, CNRS, Montpellier, fboutin@univ-montp1.fr
[b] INA, Paris, jthievre@ina.fr
[c] LIRMM – CNRS, Montpellier, mountaz@lirmm.fr

ABSTRACT

Realistic interaction networks usually present two main properties: a power-law degree distribution and a small world behavior. Few nodes are linked to many nodes and adjacent nodes are likely to share common neighbors. Moreover, graph structure usually presents a dense core that is difficult to explore with classical filtering and clustering techniques. In this paper, we propose a new filtering technique accounting for a user-focus. This technique extracts a tree-like graph with also power-law degree distribution and small world behavior. Resulting structure is easily drawn with classical force-directed drawing algorithms. It is also quickly clustered and displayed into a *multi-level silhouette tree* (*MuSi-Tree*) from any user-focus. We built a new graph filtering + clustering + drawing API and report a case study.

Keywords: filtering technique, graph clustering, focus-based approach, social network

1. INTRODUCTION TO REALISTIC INTERACTION NETWORKS

Interaction networks are information structures (graphs) including actors and relations between actors. They are used in computer science (web graph, internet network), bibliometry (citation and co-authoring graphs), sociology (relationship network), biology (protein interaction graph) or economy (sharing or communication networks). Visualizing and exploring realistic networks is challenging, since they usually present a very dense core.

Two main models were developed for realistic interaction networks: "small world" model and "scale free" model.

"Small world" model lies on two properties: the first one, called "small world phenomenon", was discovered in a famous experiment on social network[1]: relationship-distance between people is small (6 degrees of separation). Second property called "clustering property" was settled by Watts and Strogatz [2]: two connected nodes are more likely to share common neighbors than random nodes. It is measured by a global clustering coefficient C computed by the average of local clustering coefficients C(v). Considering a vertex v with N_v neighbors and E_v edges between them:

$$C(v) = \frac{2E_v}{N_v(N_v - 1)} \text{ if } N_v > 1 \text{ and } C(v) = 0 \text{ if } N_v = 0 \text{ or } 1 \quad (1)$$

Watts and Strogatz introduced a "small world" graph construction[2]. Nodes are first organized on a ring and connected with their k nearest neighbors. Then, some links are disconnected and random links are added.

Kleinberg[3] proposed another construction using three constants p, q, r. Nodes are first displayed on a 2D grid and connected with their p-neighborhood (nodes at distance almost p). Then, each node is connected to q remote nodes so that distance d between the two nodes is proportional to d^{-r}. Resulting edge is called "long-range" edge.

These two models are proved to be "small world" even though long-range edges (or random edges) are few. They present a large clustering coefficient (like lattice), a small diameter and Poisson degree distribution (like random graph). However, interaction networks have usually a power law degree distribution[4] instead of a Poisson distribution.

In 1999, Barabási and Albert[4] introduced the "scale free" model whose construction is based on a growing process. Each new node is connected to the graph with preferential attachment, so that nodes with high degree are more likely to be connected. By construction, this model presents a power law degree distribution defined by:

$$P_k = a.k^{-\beta} \text{ where } P_k \text{ is the probability to find a node of degree } k. \quad (2)$$

Visualization and Data Analysis 2006, edited by Robert F. Erbacher, Jonathan C. Roberts, Matti T. Gröhn, Katy Börner,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6060, 60600Q, © 2006 SPIE-IS&T · 0277-786X/06/$15

Scale free networks are often attached the “rich get richer” mantra. A few nodes have relations with many others. This property is very common in realistic interaction networks like web graphs or citation graphs that often present hubs. Consequently these graphs usually have a dense “core”. Unfortunately classical filtering + clustering techniques fail to organize scale free networks into clusters. Indeed, graphs are usually peeled like onions (removing “external layers”) or split up into many components without relations between them. However, none structure is extracted from the “core”.

We studied various realistic networks that present both power-law degree distribution and small world phenomenon. We propose a new filtering technique that transforms these graphs into “tree-like graphs” with also scale free and small world behavior. This technique is specific since it accounts for “filtering focus” and “filtering threshold”. So, the graph can be filtered from various perspectives, with different levels of filtering.

Resulting filtered graph is well displayed with energy-based methods[5,6] but also radial drawing techniques[7]. Moreover, filtered graph can be clustered into a “multi-level silhouette tree” (*MuSiTree* [8]) that is easily drawn and managed. Our clustering technique is fast and takes into accounts various “clustering foci”, so that changing focus provides different views of a “tree-like graph”.

So, initial graph can be filtered from various “filtering foci” and “filtering thresholds”. Moreover, a “tree-like graph” can also be organized from various “clustering foci”. We developed a specific drawing technique to explore *MuSi-Trees* using *Prefuse* API[9].

A real co-authoring network is presented in section 2 to illustrate the various techniques. Then, filtering methods are reviewed and a new focus-based filtering technique is introduced in section 3. Our filtering technique is coupled with a focus-based clustering technique, in section 4, providing a “multi-level silhouette tree”.

2. CASE STUDY: A CO-AUTHORING NETWORK

In this paper we consider a real co-authoring graph that contains 1511 nodes (authors) and 7902 co-authoring links. This graph was collected from 2000 to 2004 in the laboratory of computer science, robotic & microelectronic (LIRMM) of Montpellier.

We present (Fig. 1) a spring view[10] of the largest component of the co-authoring graph (80 % nodes). Colors represent various fields (blue edges: microelectronic, pink edges: robotic, green and yellow edges: computer science). Nodes with high degree are red when nodes with small degree are green.

Unfortunately, resulting view is cluttered since there are many interactions between authors.

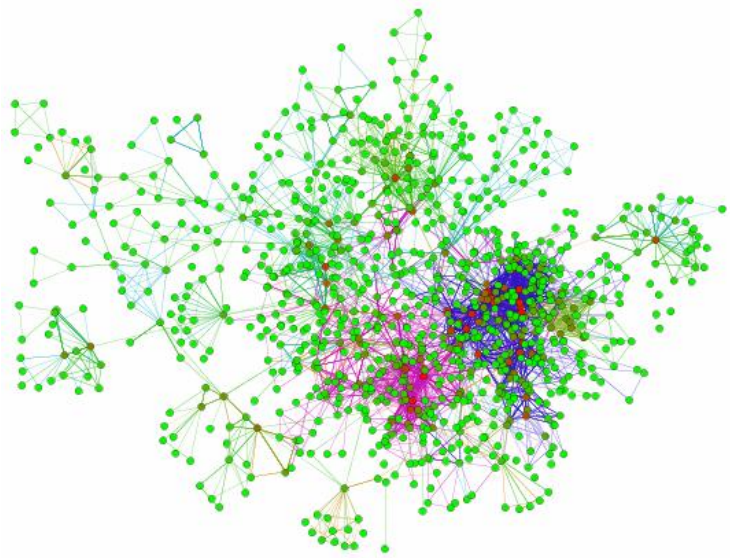


Figure 1. Co-authoring graph: 1511 nodes and 7902 edges – spring view

Co-authoring graph presents a power law degree distribution. So, the graph is said "scale free" (see Fig. 2).

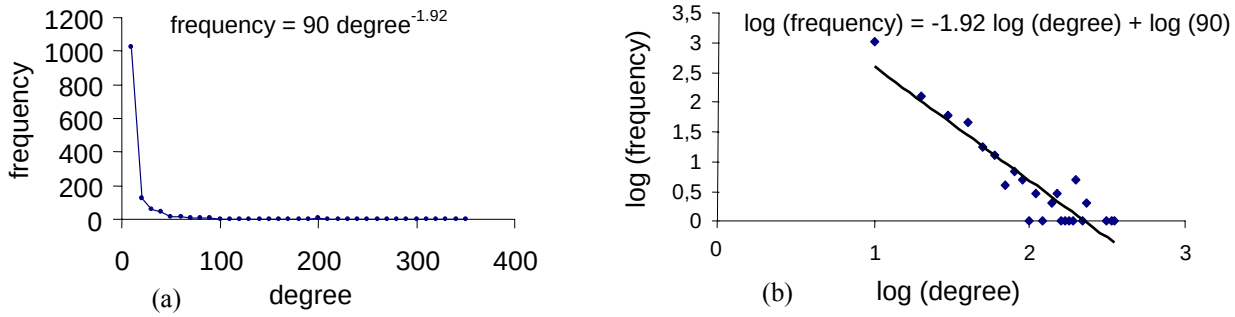


Figure 2. Power law degree distribution (a) linear scale (b) log log scale

Local clustering coefficients have also a power law distribution (see Fig. 3). Note that C(v) is sometimes above one since two authors can write many papers together (weighted graph). Resulting global clustering is 2.7 and average diameter is around 12. So, the co-authoring graph is a small world graph.

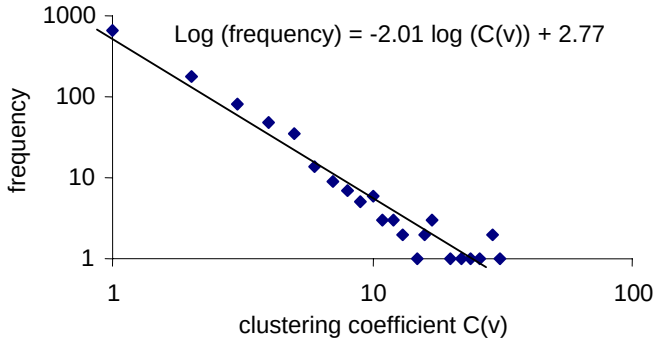


Figure 3. Power law clustering coefficient distribution – log log scale

Degree and local clustering coefficient are correlated within a power law distribution (Fig. 4). On one hand, authors with high degree (hubs) have usually low clustering coefficient. It is normal since a hub neighborhood is rarely completely related. On the other hand, many authors have a large clustering coefficient. They usually belong to a dense community without relations with other communities.

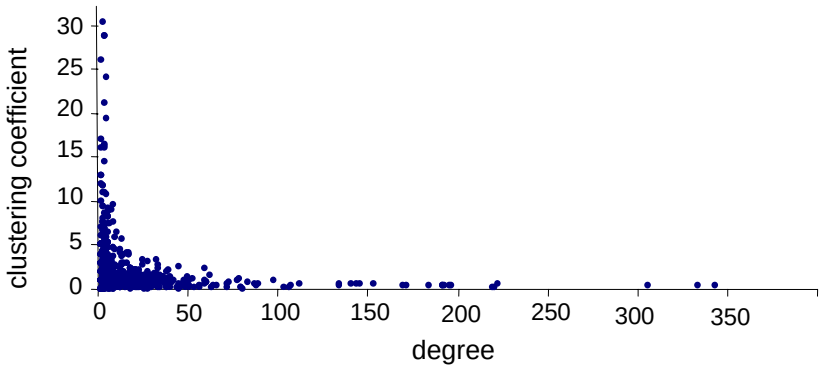


Figure 4. Correlation between degree and clustering coefficient

In next section, various filtering techniques are applied to this "scale free"+ "small world" co-authoring graph.

3. FILTERING PROCESSES

Filtering a graph consists in removing nodes or edges according to a qualitative or quantitative metric on these elements. The objective of a filtering technique is to simplify a graph in order to extract a graph structure or interesting components without losing graph properties.

In this work, we focus on structure dependant metrics and review filtering techniques in field of scale free networks with small world behavior. Then we present a new filtering technique well suited to display such graphs.

3.1. Review of Filtering Techniques

Many filtering techniques consist in removing nodes or edges using a specific metric[11]: a centrality metric (degree, closeness or betweenness centrality[12]), or a clustering metric (clustering coefficient[2] or edge force[13]).

Removing "weak" nodes with degree < 4 (Fig. 5a) or degree < 10 (Fig. 5b) simplifies the graph ("external layers" are removed). This technique fails to extract components in the dense core. However, it can be used as a preliminary filtering technique.

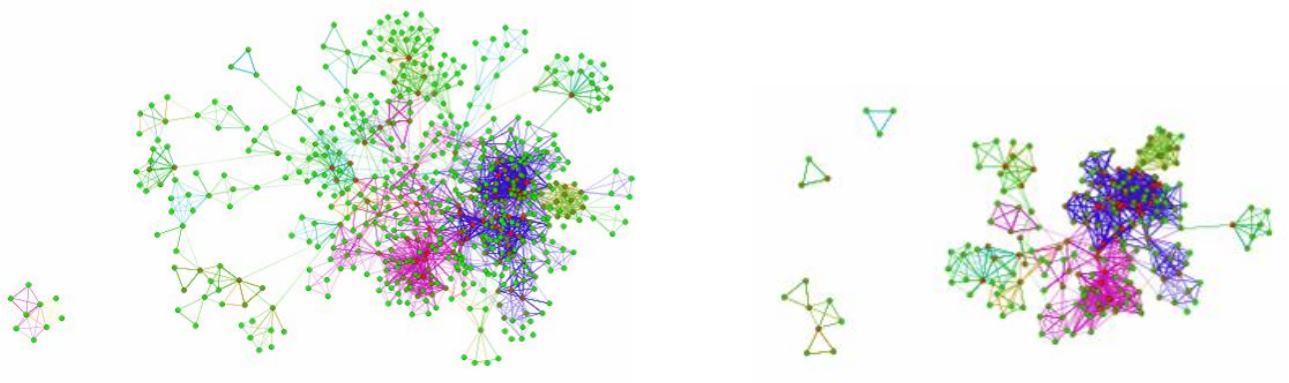


Figure 5. Removing nodes with (a) degree < 4 (b) degree < 10

Node betweenness centrality (BC[12]) is a centrality index defined by the fraction of shortest paths between node pairs passing through a node of interest. Filtering nodes with low betweenness centrality peels the graph (removing "secondary nodes"). Unfortunately, resulting graph remains cluttered (Fig. 6).

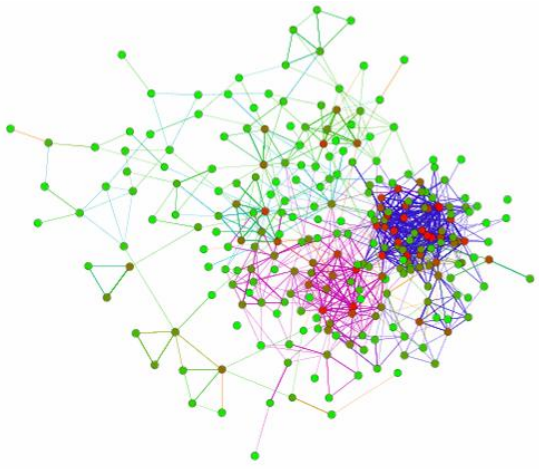


Figure 6. Removing nodes with betweenness centrality null (263 nodes, 3516 edges)

Another filtering approach consists in removing nodes with small local clustering coefficient[2] in order to extract patterns. Unfortunately, this technique often removes hubs and articulation nodes so that resulting graph is split up into many connected components without relations between each other (Fig. 7).

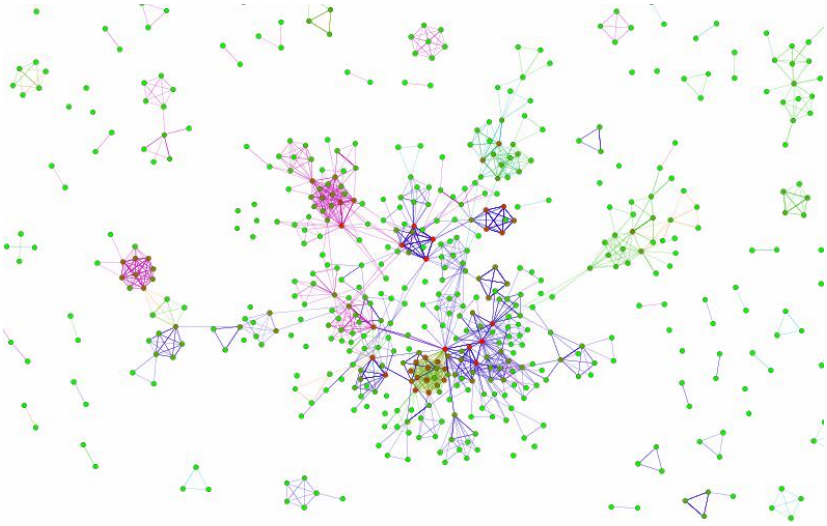


Figure 7. Removing nodes with clustering coefficient < 0.5 (910 nodes, 2845 edges)

We can also use an edge filtering technique to simplify the graph, without removing any node. A filtering technique based on *edge strength* computing[13] is applied on co-authoring graph (Fig. 8). *Edge strength* metric generalize clustering coefficient to edges. An edge (u, v) is said weak if u and v have "almost" no common neighbors. Weak edges with strength < 0.5 are removed. Even though, some small components are disconnected, the "core" remains very dense.

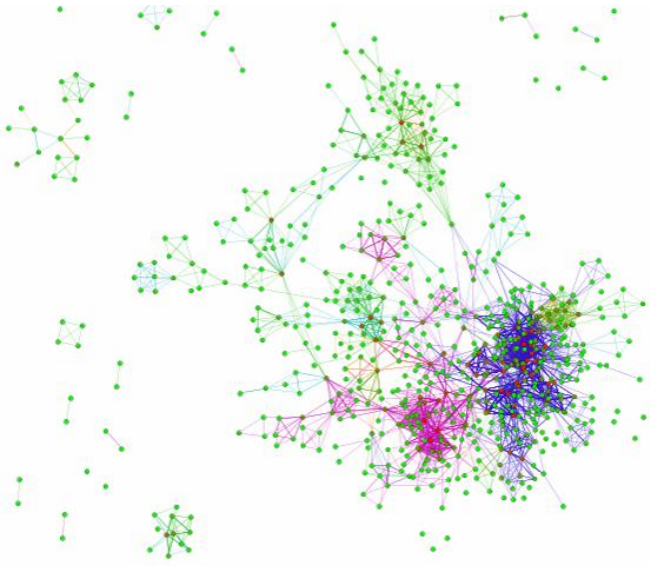


Figure 8. Removing edges with strength < 0.5 (1511 nodes, 6828 edges)

Previous techniques usually fail to organize a "scale free" graph with dense "core". Networks are either peeled like onions (removing "external layers" or "secondary nodes") or split up into many small components (without relations between them). We propose, in next section, a new filtering technique to simplify and structure a scale free graph.

3.2. A New Focus-based Filtering Technique

A filtering technique aims to remove nodes or edges in order to extract a natural structure or some interesting components. We argued that splitting dense "core" of a scale free graph with classical filtering techniques is challenging.

In this section, we introduce a new edge filtering technique that accounts for a "filtering focus". This technique is applied to any connected component. It consists in skeleton extraction and dense components extraction.

Skeleton Extraction:

We propose a spanning tree extraction from any "filtering focus" with preferential attachment. Spanning tree is built iteratively. Let denote $T_n = (V_n, E_n)$ the spanning tree at step n:

- V_0 is defined by the "filtering focus" (tree root).
- In order to build V_{n+1} from V_n we add the new node with highest degree related (in the graph) to any node in V_n. Then this node is connected (in the tree) to its neighbor in V_n with highest degree.

Thus, resulting maximal spanning tree is also "scale free" since nodes are related to their neighbor with highest degree. Another technique is introduced[14] that does not account for any user-focus.

We apply our spanning tree extraction technique to our co-authoring graph with a first user-focus: for instance, node of highest degree (Fig. 9a). Spanning tree has power low degree distribution (Fig. 9b) like initial graph (Fig. 2). Moreover, power law coefficient 1.96 is closed to initial coefficient 1.92.

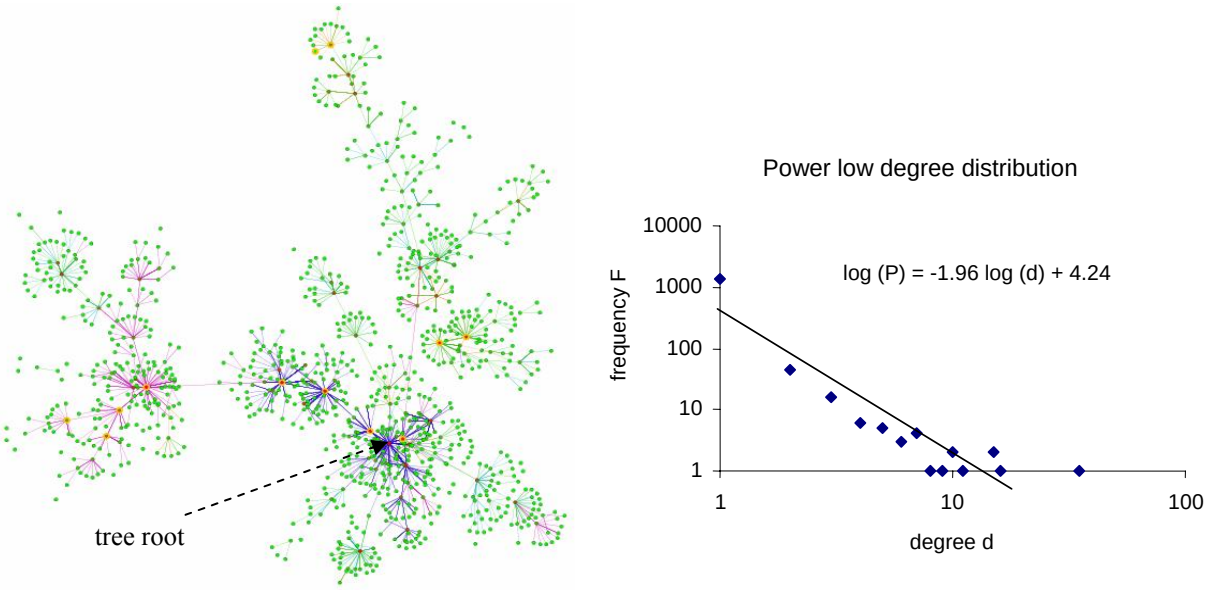


Figure 9. Spanning tree extraction – (a) spring view (b) power law degree distribution

Natural communities (computer science, robotic and microelectronic laboratories) are well separated (but not disconnected). However, too many edges are removed so that clustering coefficient is null.

Next, we propose to keep some non-tree edges to extract dense components.

Dense Components Extraction:

Previous skeleton is scale free, like initial graph, but it does not present clustering. To keep a large clustering coefficient, we propose a new technique that consists in adding some "short edges" to the spanning tree. A short edge is an edge that connects two nodes at short distance in the tree.

The algorithm computes, for each non-tree edge, the length of the shortest path between its two nodes in the tree. It consists in finding the lowest common ancestor in the tree. Then, "long edges" are removed from initial graph.

The structure defined by tree skeleton and "short" edges is called "tree-like graph". It has various properties:

- Its diameter is close to spanning tree diameter but larger than initial graph.
- Its clustering coefficient is close to initial graph clustering coefficient.
- If initial graph is scale free, "tree-like graph" is also scale free (like spanning tree).

So, removing long edges in a small world + scale free graph provides a "tree-like" small world + scale free graph. Moreover, components are likely to be well separated but not disconnected.

The algorithm is applied to the co-authoring graph (Fig. 10) with two thresholds (6 and 4). The second filtered graph (Fig. 10b) is well organized into dense sets of nodes. Resulting structure looks like spanning tree (Fig. 9a).

In Fig. 11, we propose a graph organization from another focus: "Guy Mélançon". Microelectronic and robotic domains were very closed in Fig. 10. However, they seem well separated with the new focus.

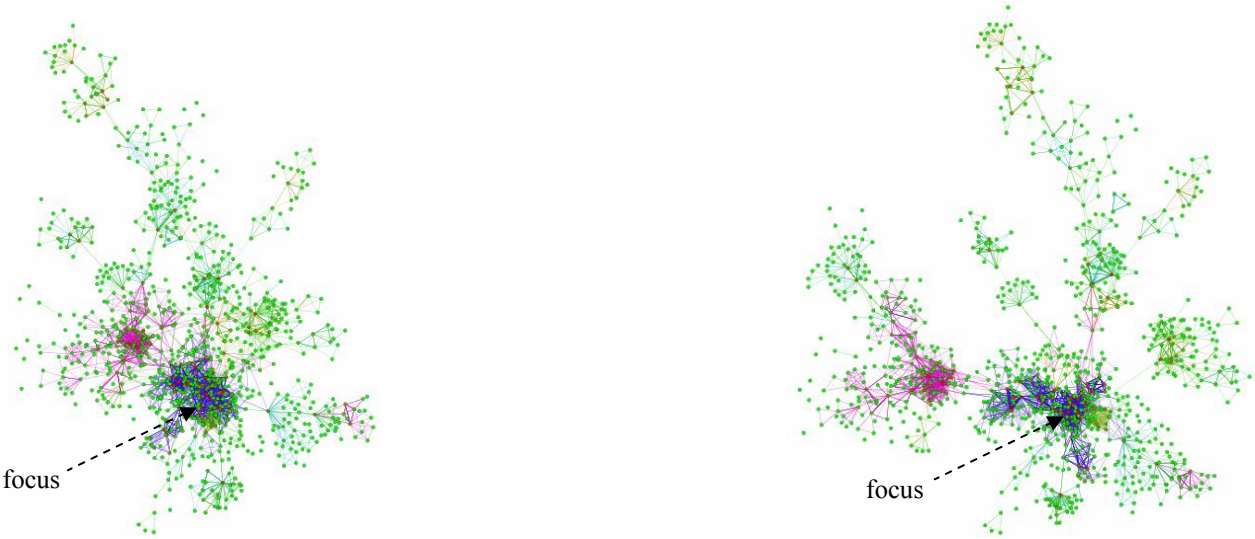


Figure 10. "Long edges" removal with (a) length > 6 and (b) length > 4

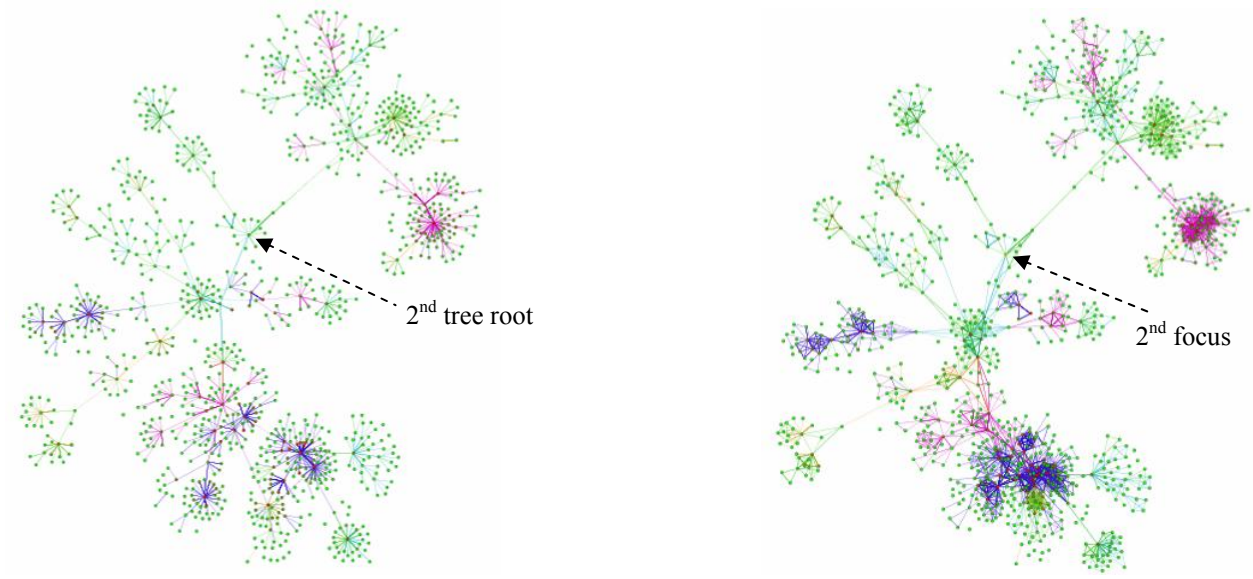


Figure 11. Second focus "Guy Mélançon" (a) spanning tree (b) removing edges with length > 4

Consequently, scale free graphs with small world phenomenon can be organized into "tree-like graphs" from various perspectives. Resulting filtered graph is easily drawn with a force-directed drawing technique. It looks like a tree of "dense components" organized around the "filtering focus".

Next, we describe a focus-based clustering technique that automatically identifies these dense components from any "clustering focus".

4. CLUSTERING PROCESS

Graph clustering techniques consist in organizing nodes into clusters so that relations inside clusters are "stronger" than relations between clusters[15]. We review three types of techniques: geometric, metric-based and structural techniques (section 4.1). Then we present a focus-based clustering technique (section 4.2) applied to "tree-like graphs" (section 4.3).

4.1. A Brief Review of Clustering Techniques

Geometric clustering techniques, introduced in data mining[16], suppose that graphs are defined in a vector space (geographic graphs in 2D or 3D vector space; web graphs with pages as vectors in term-space[17]; graphs displayed in 2D space with force-directed drawing algorithm [10]). Their main objective is usually to define an optimal surface (sphere or hyperplane[18]) that splits the graph into two components, minimizing edge cuts. This technique is applied recursively. K-means is another technique that organizes nodes iteratively minimizing inertia around centroids[19] (barycenters).

Metric-based clustering techniques depend on structural or geometric metric. The most popular is the hierarchical clustering technique that proceeds iteratively to merge closest clusters according to a metric[19]. Resulting clustering hierarchy, called dendrogram, is cut to produce an optimal partition.

Structural clustering techniques are specific to graph structure. An optimization technique (KL algorithm[20]) consists in exchanging nodes between clusters in order to minimize edge cuts. Other techniques are based on random walks[15,18], flow simulation[21] or spectral analysis[15,18].

Previous techniques do not account for user-focus. However, organizing a graph from a user perspective may be attractive. Next, we present a focus-based clustering technique that is well suited to organize "tree-like graphs"[8].

4.2. A focus-based Clustering Technique

We recently developed a focus-based clustering technique[8] that transforms any connected graph into a multi-scale structure called "multi-level silhouette tree" (*MuSi-Tree*). It is a tree of components (called silhouettes). Each silhouette is also a tree of nested silhouettes. This structure is computed in linear time from any "clustering focus".

MuSi-Tree construction is briefly recalled. Let consider a connected graph G and a "clustering focus":

- G_k is defined as the sub-graph of G containing nodes at distance at most k from the focus.
- G_k can be organized into a tree of biconnected components (silhouettes). To avoid silhouettes overlapping, each articulation node is set in the first component encountered from the focus and remove from the others.

By construction, any silhouette of G_k belongs to a silhouette tree of G_{k+1}. Consequently, resulting compound structure is hierarchically clustered.

The algorithm is illustrated with a small graph G (Fig. 12). G and G_2 are organized into biconnected components from focus 1 (Fig. 12a). Graph G presents 4 articulation nodes (1, 20, 24 and 30). G_2 has also 4 articulation nodes (1, 5, 10 and 12). Articulation nodes are set in the first encountered silhouette (Fig. 12b).

Some techniques were developed to visualize and interact with hierarchical structures[22-25]. We developed an API devoted to *MuSi-Tree* visualization and interaction. It is based on radial graph layout[7,9]. Nested silhouettes are painted with a same color (Fig. 12b). We decided to draw links between nodes instead of links between silhouettes.

Changing focus (Fig. 13) dynamically reorganizes the view without changing global silhouettes structure (except articulation nodes position). It produces various perspectives of a same graph.

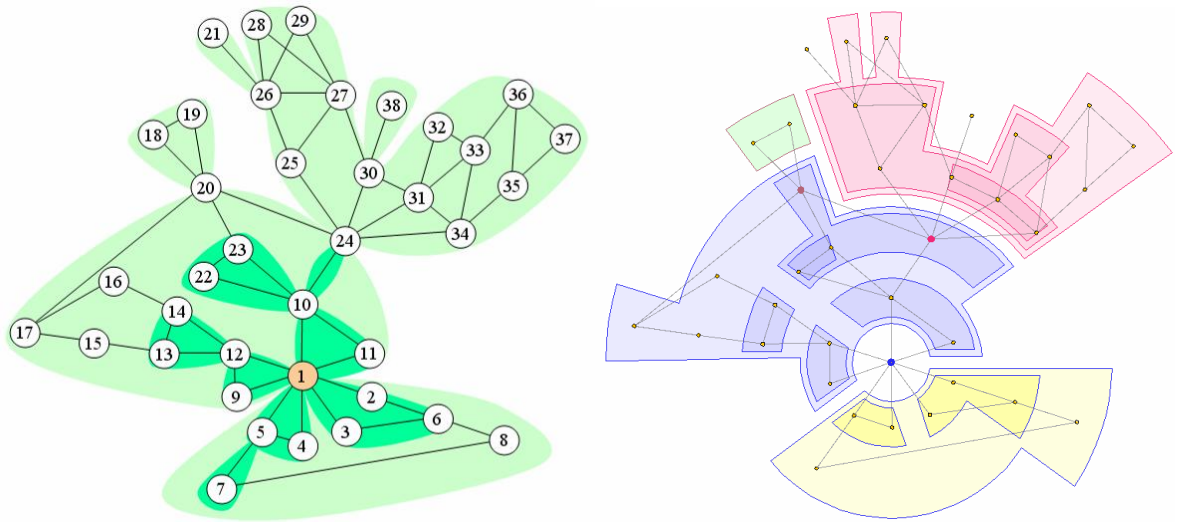


Figure 12. (a) biconnected components organization (b) *MuSi Tree* from focus 1

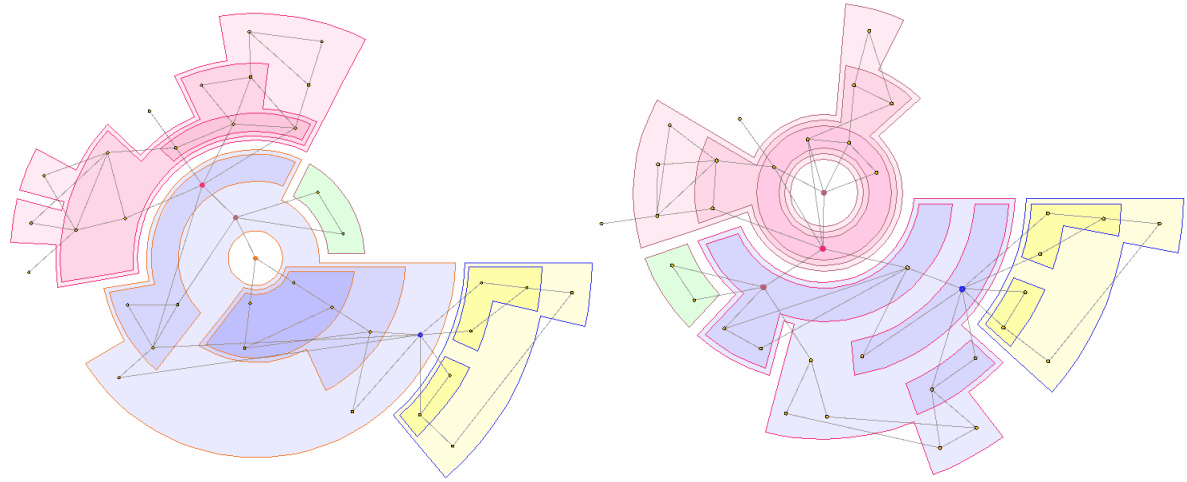


Figure 13. (a) *MuSi Tree* from focus 17 (b) *MuSi Tree* from focus 31

We show in next section that this clustering + drawing technique is likely to organize "tree-like graphs".

4.3. "Tree-like Graph" Clustering

We introduced in section 3.2 a focus-based filtering technique that organizes a scale free graph (with dense "core") into a specific structure called "tree-like graph". Resulting filtered graph is easily drawn with force-directed drawing algorithm.

Moreover, the focus-based clustering technique [8] described in section 4.2 is well suited to organize "tree-like graph". Indeed, a "tree-like graph" is likely to have many articulation nodes between components (Fig.s 10b, 11b).

Co-authoring Graph: clustering + drawing technique is applied to the filtered graph (Fig. 11b) from two clustering foci (Fig. 14). The first "clustering focus" is "Guy Mélançon" (from computer science laboratory). The second one is: "Serge Bernard" (from microelectronic laboratory). Resulting views represent the same filtered graph from two perspectives. Note that "filtering focus" and "clustering focus" may be different.

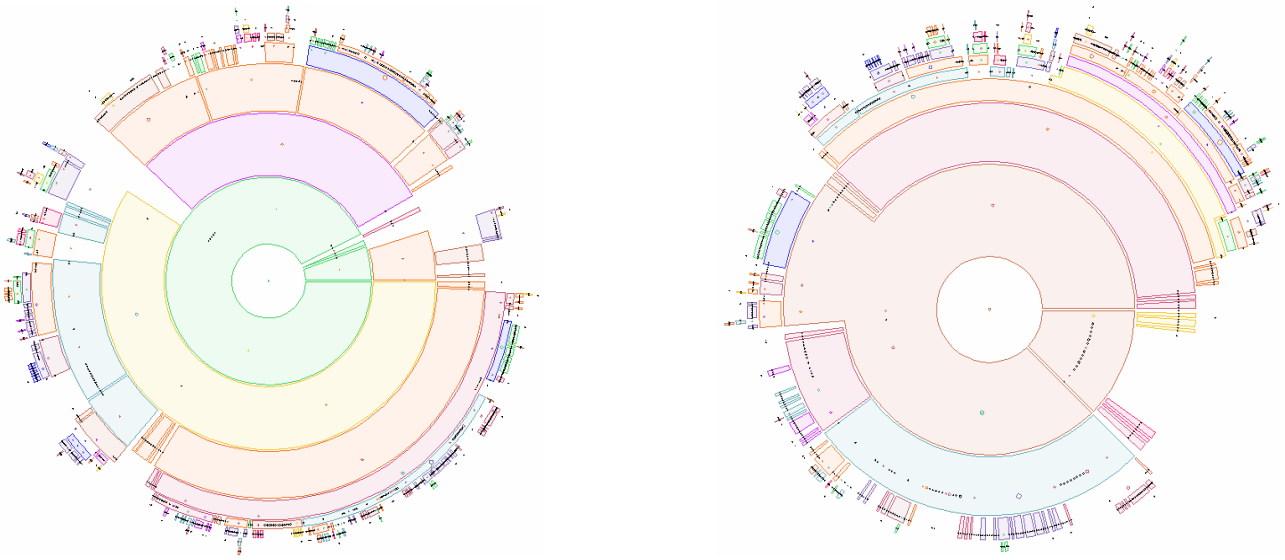


Figure 14. *MuSi Tree* from two foci (a) "Guy Mélançon": computer science (b) "Serge Bernard": microelectronic

In order to analyze filtering + clustering effectiveness, we propose to study the "dual" graph of co-authoring graph:

Dual Graph: We show (Fig. 15a) a spring view of co-authoring dual graph where nodes represent papers (instead of authors). The dual graph is very dense since it contains 2211 nodes and 53722 edges. Nodes color represents specific domains (robotic, microelectronic and computer science). Edges color represents authors.

We apply our filtering technique on dual graph from a paper focus called "*Software component capture using graph clustering"* [13] including "short edges" of length at most 4. Resulting "tree-like graph" is drawn in Fig. 15b (2211 nodes and 47083 edges). Papers are well separated in domains.

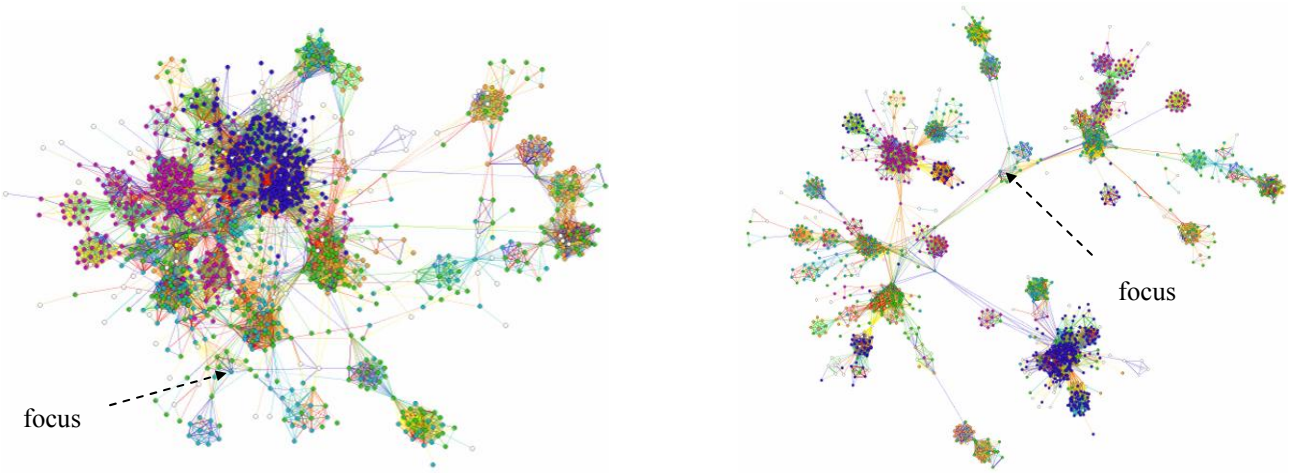
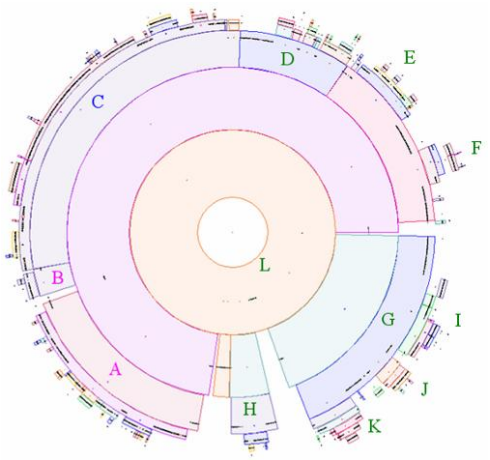


Figure 15. *(a) Initial dual graph (b)* removing edges with length > 4

We first apply our clustering technique using focus[13] (Fig. 16). Silhouettes are manually labeled with 12 domains. The technique could be improved including automatic labeling with articulation node information or text frequency.

Then, we apply our technique from the same focus, removing edges of length > 3. Resulting *MuSi-Tree* (Fig. 17) looks like previous *MuSi-Tree* (Fig. 16) where some components are grouped.



Papers organization in domains:

A: synthesis of dynamical complex systems (robotic)
B: management and supervision of complex systems (robotic)
C: microelectronic
D graph theory
E: machine learning
F: object and agents
G: data base
H: arithmetic and combinatory
I: navigation, perception and control of mobile vehicles (robotic)
J: human computer interface
K: language treatment
L: graph drawing

Figure 16. *MuSi-Tree* of filtered dual graph - removing edges with length > 4

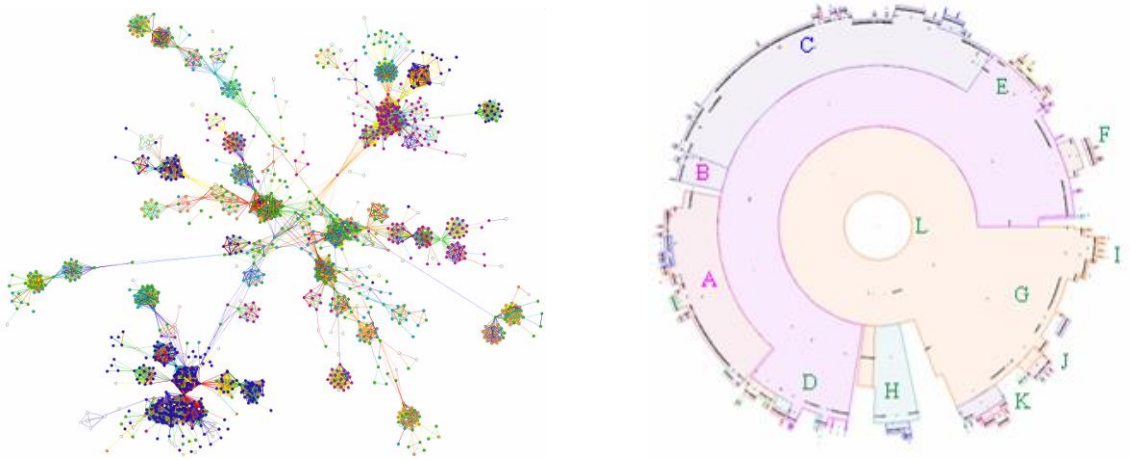


Figure 17. (a) Filtered dual graph - removing edges with length > 3 (b) *MuSi-Tree*

5. CONCLUSION

Main realistic interaction networks present power-law degree distribution and small world behavior. However, classical filtering and clustering techniques usually fail to split up the dense core into an organized structure.

We developed a new filtering method that extracts both graph skeleton and dense components from scale free graphs. Resulting filtered graphs are also scale free with small world behavior, but their "tree-like" structure is easier explored. Moreover, filtering technique depends on "filtering focus" so that graph organization accounts for user perspective.

Our filtering technique is fast and easily applied iteratively for various user foci and thresholds. Energy models layout algorithms can be used to display resulting "tree-like" structures. Our technique can be also coupled with a clustering technique. We introduced[8] a clustering technique especially devoted to "tree-like graphs" that accounts for a user "clustering focus". The "tree-like graph" is organized into a tree of nested components called *MuSi-Tree* ("multi-level silhouette tree") around "clustering focus".

Initial case study on co-authoring graph (and its dual graph) within the members of a research lab provided good insights on what the main structure of the graph was. Our filtering + clustering technique is well suited to display actual collaborations between researchers as well as main trends and strong groups. In the future, we plan to use both analytical criteria[26] and controlled experiments to provide comparative evaluation of our technique.

REFERENCES

1. Milgram, S., 'The Small World Problem': *Psychology Today*, p. 60-67, 1967
2. Watts, D. J., and S. H. Strogatz, 'Collective dynamics of "small-world" networks': *Nature*, v. **393**, p. 440-442, 1998
3. Kleinberg, J., 'The Small-World Phenomenon: An Algorithmic Perspective': *Cornell Computer Science Technical Report*, v. **99-1776**, 1999
4. Barabási, A. L., and R. Albert, 'Emergence of scaling in random networks': *Science*, v. **286**, p. 509-512, 1999
5. Battista, G. D., P. Eades, R. Tamassia, and I. G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*, 432 p, Prentice Hall, 1999
6. Noack, A., 'An energy model for visual graph clustering': *Graph Drawing*, p. 425-436. 2003
7. Yee, K. P., D. Fisher, R. Dhamija, and M. A. Hearst, 'Animated Exploration of Dynamic Graphs with Radial Layout': *Information Visualization*, p. 43-50. 2001
8. Boutin, F., J. Thièvre, and M. Hascoët, 'Multilevel Compound Tree - Construction Visualization and Interaction': *Interact*. 2005
9. Heer, J., S. K. Card, and J. A. Landay, 'Prefuse: a toolkit for interactive information visualization': *CHI, Human Factors in Computing Systems*. 2005
10. Fruchterman, T. M. J., and E. M. Reingold, 'Graph drawing by force-directed placement.' *Software – Practice and Experience*, v. **21**, p. 1129-1164, 1991
11. Huang, X., P. Eades, and W. Lai, 'A Framework of Filtering, Clustering and Dynamic Layout Graphs for Visualization': *ACSC 2005*, p. 87-96, 2005
12. Brandes, U., 'A faster algorithm for betweenness centrality': *Journal of Mathematical Sociology* v. **25**, p. 163-177, 2001
13. Chiricota, Y., F. Jourdan, and G. Mélançon, 'Software component capture using graph clustering': *11th International Workshop on Program Comprehension*, p. 217-226, 2003
14. Kim, D. H., J. D. Noh, and H. Jeong, 'Scale-free trees: the skeletons of complex networks': *Physical Review*, v. **E 70**, 2004
15. Alpert, C. J., and A. B. Kahng, 'Recent Developments in Netlist Partitioning: A Survey, Integration': *VLSI Journal*, v. **19**, p. 1-81, 1995
16. Jain, A. K., M. N. Murty, and P. J. Flynn, 'Data clustering: A Review': *ACM Computing Surveys*, v. **31**, p. 264-323, 1999
17. Han, E. H., and G. Karypis, 'Centroid-Based Document Classification: Analysis and Experimental Results': *4th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD)*, p. 427-431. 2000
18. Chamberlain, B. L., 1998, Graph Partitioning Algorithms for Distributing Workloads of Parallel Computations, Washington, University.
19. Karypis, G., E.-H. Han, and V. Kumar, 'CHAMELEON: A hierarchical clustering algorithm using dynamic modeling': *IEEE Computer*, v. **32**, p. 68-75, 1999
20. Kernighan, B. W., and S. Lin, 'An efficient heuristic procedure for partitioning graphs': *Bell System Technical Journal*, v. **49**, p. 291-308, 1970
21. van Dongen, S., 2000, Performance criteria for graph clustering and Markov cluster experiments, Amsterdam, National Research Institute for Mathematics and Computer Science in the Netherlands.
22. Sugiyama, K., S. Tagawa, and M. Toda, 'Methods for Visual Understanding of Hierarchical System Structures': *IEEE Transactions on Systems Man and Cybernetics*, v. **11**, p. 109-125, 1981
23. Sugiyama, K., and V. Misue, 'Visualization of structural information: automatic drawing of compound graphs': *IEEE Trans. System Man Cybernetics*, v. **21**, p. 876-892, 1991
24. Eades, P., 'Multilevel Visualization of Clustered Graphs': *Graph Drawing*, p. 101-112. 1996
25. van Ham, F., and J. J. van Wijk, 'Interactive Visualization of Small World Graphs': *InfoVis*, p. 199-206. 2004
26. Boutin, F., and M. Hascoët, 'Cluster Validity Indices for Graph Partitioning': *Information Visualisation*. 2004