Visualization and analysis of large graphs

Eloïse Loubier
Ph.D.
Institut de Recherche en Informatique de Toulouse
118 route de Narbonne
31062 Toulouse Cedex 9
O5 61 55 67 81, France
loubier@irit.fr

Wahiba Bahsoun
Institut de Recherche en Informatique de Toulouse
118 route de Narbonne
31062 Toulouse Cedex 9
O5 61 55 69 45, France
wahiba.bahsoun@irit.fr

Bernard Dousset
Institut de Recherche en Informatique de Toulouse
118 route de Narbonne
31062 Toulouse Cedex 9
O5 61 55 67 81, France
dousset@irit.fr

ABSTRACT
In Knowledge engineering, synthesized information has often an evolving and relational form. Information representation using graphs may ease data interpretation for non-expert users. However this graph may be complex and simplifications are useful in order to ease analysis. In this article, we present VisuGraph, a powerful tool for graph drawing. This tool gives the possibility to reduce large graph by two techniques: the Markov CLustering algorithm (MCL) application and the global graph division in time-sliced visualizations in order to specify and to simplify temporal analysis.

Categories and Subject Descriptors
I.3.3 [**Computer Graphics**]: Methodology and techniques – *Ergonomics, Graphics data structures and data types, interaction techniques; H.5 [User Interfaces] Graphical User Interface.*

General Terms
Algorithms, measurement, documentation, performance, design, reliability, experimentation, human factors, standardization.

Keywords
large graph, clustering, dynamic graph visualization, temporal graph, collaboration graph, reduced graph.

1. INTRODUCTION
In knowledge engineering, synthetic information often takes a relational form based on the links between actors, semantic networks, alliances, fusions, acquisitions, collaborations or co-occurrences. The representation of this type of information in the form of optimized graph is analyzable by non-experts in data analysis. This graph structure shows actor characteristics and relations between them. One may see this kind of visualization in several domains: data processing (Web graph, Internet network); bibliometry (quotations graph, co-authors graph); sociology (personal or professional collaboration graph); biology (graph of protein interactions); and geography or economy (exchange or communication networks).

More generally, networks are the backbone of complex systems that consists of many components (actors, often called “nodes”, can be persons, teams, organizations, concepts...). Understanding network structures would greatly help understanding the way components interact with each other as well as the intrinsic properties of complex systems.

The visualizations in existing systems do not satisfactorily answer the basic dilemma of being readable not only for the global structure of the network but also for detailed analyses of local communities.

Especially, evolving networks display a high degree of complexity, due to the inherent wiring entanglement occurring during their growth. The effectiveness of network visualization techniques also depends on network size. Two approaches may be chosen: firstly to simplify the network in a reduced graph where each node represents an actor group; secondly to explode the global graph into several subgraphs. Clustering is also a solution to simplify the network structure. These partitioning techniques are numerous, but the Markov clustering algorithm (MCL) has the property to take into account the transitive aspect of the graph structure. This method is added to the graph drawing tool named VisuGraph [1]. With Markov Clustering it is possible to obtain homogeneous classes. The result is a reduced graph where each class can be represented by its most significant node. This clustering technique based on the spectral analysis takes into account the graph topology, but it requires a number of classes. We modify it to be able to randomise the number of classes (increasing or reducing the number). In spite of this network structure, the MCL algorithm is not sufficient and a partial study is essential for evolving data. The total graph can be divided in several period graphs if the temporal dimension is taken into account.

Global period division in time-slices simplifies the general analysis, which can be complex and lead to analysis errors. Indeed, a presumed related class on a global graph may never have this characteristic if each period is studied separately. With these two solutions each class may first be analyzed separately on the reduced graph (small-scale model of the global graph), and then on a specific time-slice.
Furthermore, this work shows the usefulness of taking into account network structure as a number of measures within the bounded context of temporal subsets. Differences in network topology and properties within each period should have predictive consequences and the relative duration of clusters.

The remainder of this paper is structured as follows: in section 2, we present VisuGraph, the graph drawing tool used in this study; in section 3, we explain how to simplify a complex large graph by the means of the MCL or by temporal analysis. Lastly, we conclude and we present the prospects for future work.

2. THE VISUGRAPH TOOLKIT

Network maps that facilitate studying the topology of such large complex networks were first developed in different disciplines in the late 1990s. As Dorogotsev and Mendes [6] wrote: "The first experimental data, mostly for the simplest structural characteristics of the communication networks, were obtained in 1997-1999." Recent papers [7, 8, 9], contain detailed summaries of the ongoing work in the field of complex networks.
In this study, graphs with nodes and links are drawn to address primary questions about network structure and dynamics by using the VisuGraph network display software for network visualization and analysis. VisuGraph [1] is a module of the strategic watch platform Tetralogie [1]. Knowledge is extracted from corpus and synthesized in the form of co-occurrences matrices by Tetralogie processing. VisuGraph then represents this data in the form of graph.

A network is a set of vertices (nodes) connected by edges (links), representing individuals and the interactions among them respectively. This longitudinal study focuses here primarily on networks with vertices differentiated only by actors' name [1] and a single type of edge, the collaboration.
A weighted spring embedder was employed to assign node locations, using an algorithm called Force-Directed-Placement (FDP) and developed following the works of Eades [3], Fruchterman and Reingold [4], and Frick et al. [5]. Spring embedders are based on the notion that the nodes may be thought of as pulling and pushing one another. Nodes representing actors who are close will pull on each other, while those who are distant will push one another apart. The algorithm seeks to find an optimum in which there is minimal stress on the springs connecting the whole set of nodes.

In order to simplify large complex graphs, we propose to reduce graphs by the mean of clustering but also to divide time in several periods. For this study, the analyses are based on authors of conference proceedings papers from the 4-years period between 2003 and 2006.

3. REDUCED GRAPH

In order to simplify data analysis and graph structure, it is interesting to study different data classes. Among the work carried out on graph partitioning, [10, 11, 12] several works are based on spectral approaches whereas the algorithms of the MONGREL family [13] are based on multi levels partitioning. The method of partitioning used in VisuGraph is inspired of Markov Clustering [2] where characteristics were modified in order to be able to influence the number of classes [14].

The MCL algorithm simulates flows using (alternating) two simple algebraic operations on matrices. Its formulation is simple: there are no high-level procedural instructions for assembling, joining, or splitting of groups. Cluster structure is bootstrapped via a flow process that is inherently affected by any cluster structure present. The first operation used by MCL is expansion, which correspond with standard matrix multiplication. The second is inflation, which is mathematically speaking a Hadamard power followed by a diagonal scaling. Inflation models the contraction of flow; it becomes thicker in regions of higher current and thinner in regions of lower current. The MCL process causes flows to spread out within natural clusters and evaporate between different clusters.

The MCL algorithm is very fast, very scalable, and has a number of attractive properties resulting high-quality clusterings [15].
This method offers a macroscopic approach with the possibility to return to a complete structure, as seen in figure 1.

Within the large graph, data which are joined by strong links must be grouped in homogeneous classes. As a result, we obtain a class graph, where each class appears in the form of a characteristic node with a specific color. Links between nodes are assimilated to links between classes, as shown in figure 1. Node names can be visualized in a short or in a long way [1], as illustrated in the reduced graph (middle in figure 1). We prefer to hide nodes names for the other graphs in order to simplify visualization. Double click on a node opens a window and displaying information from the corpus about the actor appear [1] (**** on the figure 1).

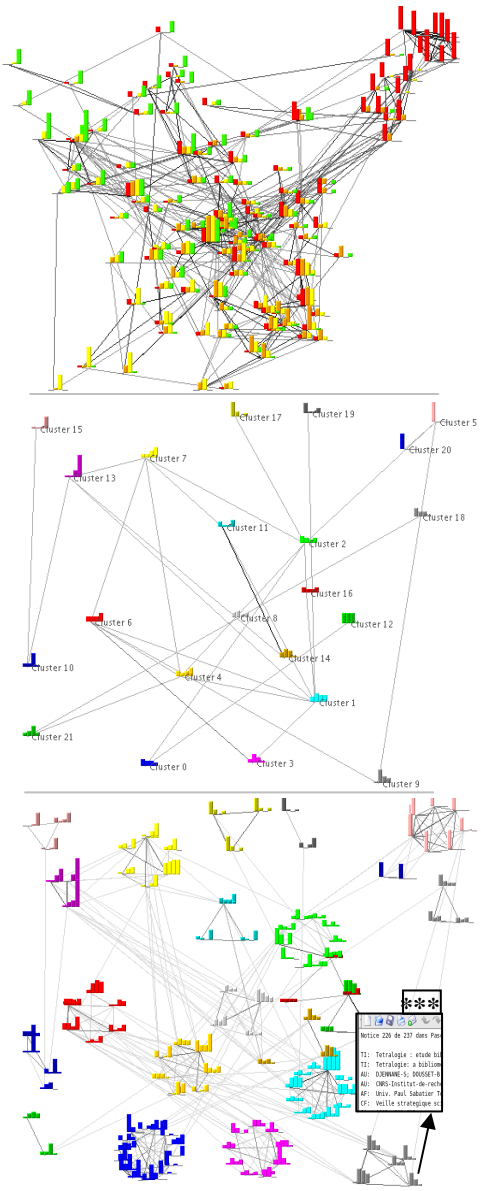

Figure 1: Simplification of a complex graph (the first graph) by the mean of the MCL method (the reduced graph is in the middle; the last graph is a global graph where one node by class is fixed).

The return to the global graph is obtained by fixing a representative node by class (the most important node of each class, which has the highest metric value). The other nodes are distributed on a circle centered on the representative node. Thus an intra-class visualization is obtained. It is possible to target the analysis and to study one or more classes. With this functionality, VisuGraph offers the possibility, by a single click on the chosen class to view one or more classes in separate windows (figure 3). VisuGraph thus makes it possible to analyze edges between classes but also to visualize the internal structure of each class.

4. ANALYSIS BY TIME-SLICES

4.1 Principle

Simplification by the MCL method is not sufficient. If a reduced graph is too complex, it is really difficult to analyze the graph structure (collaboration, alliance …) by the links between the different actors.

When the analysis is based on temporal data, it is really important to notice that the global graph could be the origin of wrong interpretations. This mistake is based on cumulative information. In fact, all temporal data are represented on the same graph and give information for all considered periods. If analyzed data are not cumulative, the interpretation in a global graph could be false and the mistake may be discovered in period graphs. As a consequence, we propose to divide time-slices in homogeneous periods. Each period is visualized in the form of graph. It is easier to analyze each time-slice and to compare them. The study is done by partial analysis, taking temporal dimension into account. For a given time period we define the collaboration graph to be a simple, undirected, node-weighted and edge-weighted graph. Vertices represent unique authors and there is an edge between two vertices if the respective authors have collaborated on a research paper.

We would like to extract information from a collaboration graph so that time-sliced visualizations gives us a better understanding of issues such as the productivity of authors, the degree of collaboration between authors, and the evolution of collaboration patterns through time.

4.2 Temporal placement

In order to take the temporal dimension into account, two steps are required. In a first step, data are visualized as a global graph, all concerned periods. In a second step, each time-slice is visualized on a period graph. Temporal data must be placed in a strategic position, which reveal their temporal characteristic. For this reason, time-slices are associated to invisible virtual nodes (reference marks) which are placed in a chronological order on the window perimeter.

Each node has an initial temporal position, calculated in function of the reference marks and its presence during each period. The more a node has an important metric value for one time-slice, the more it is situated near the corresponding temporal mark [16].

We develop a function in order to avoid nodes superposition if they belong to the same periods. Each node is joined initially with a strong link to temporal reference marks (“RMx” on figure 2) if it belong to the considered period. With temporal reference marks user can see on the global graph the tendencies and can find temporal data evolution. Positions indicate temporal data properties.

4.3 Temporal graph visualization

Dividing the global graph in period graphs makes it possible to establish a closer connection between the dynamics of actors individual, collective, alliance networks structure and evolution. Our work is based on two techniques: macro and micro analysis of reduced graph, taking temporal dimension into account.

- Macro analysis is based on the evolution of the different classes between the different time-slices (see figure 2).

- Micro analysis occurs in a specific class evolution analysis (see figure 3).

To study the growth of the evolving network, we need explicit temporal and spatial boundaries. For this study, four periods are considered, so we place one virtual reference mark on each window corner, in a chronological order. The MCL algorithm is applied and we obtain the global graph as in figure 1. Each node which represents a class is placed according to its temporal coordinates (see section 4.2.).

Central classes are the most persistent and the classes situated in window periphery are considered as changing classes. If we compare each time-sliced graph, we can observe the evolution of the links between classes. During the first and the third periods, the classes are not strongly linked. In opposition, the second and the last periods reveal the apparition of news links between them (particularly for the fourth period). In this way, the fourth time-slice appears as an important change in the global graph structure.

Class “1” (top part of figure 2) appears near the first reference mark, so we can believe that this class is more important for the first time, and less for each others. If we specify our analysis, we can compare each period graph. The first class appears just during the first period. The actors of this class are not persistent actors and disappear for each other period analyzed.

Class “2” (top part of figure 2) appears near the first and the second reference marks. Based on position of the class, class 2 belongs only to the first and the second periods. This information is confirmed if we compare each period graph and it is right to say that this class is not emergent for all concerned periods.

Class “3” (top part of figure 2) appears next to the fourth reference mark. Comparisons between time-sliced graphs confirm that this class can be seen as an emergent class. Actually, during each period this class grows and its structure is heavier for the fourth period. We can imagine it will continue to grow during futures periods.

Class “4” (top part of figure 2) is situated in the middle frame. By this strategic position, this class is persistent for the four periods. The time-sliced graphs confirm this position. This class represent some of the main actors of the graph structure. Present for each period, its structure changes are hardly visible. In this context, a micro analysis is necessarily. In figure 3, we propose to detail this class structure for each of the four periods.

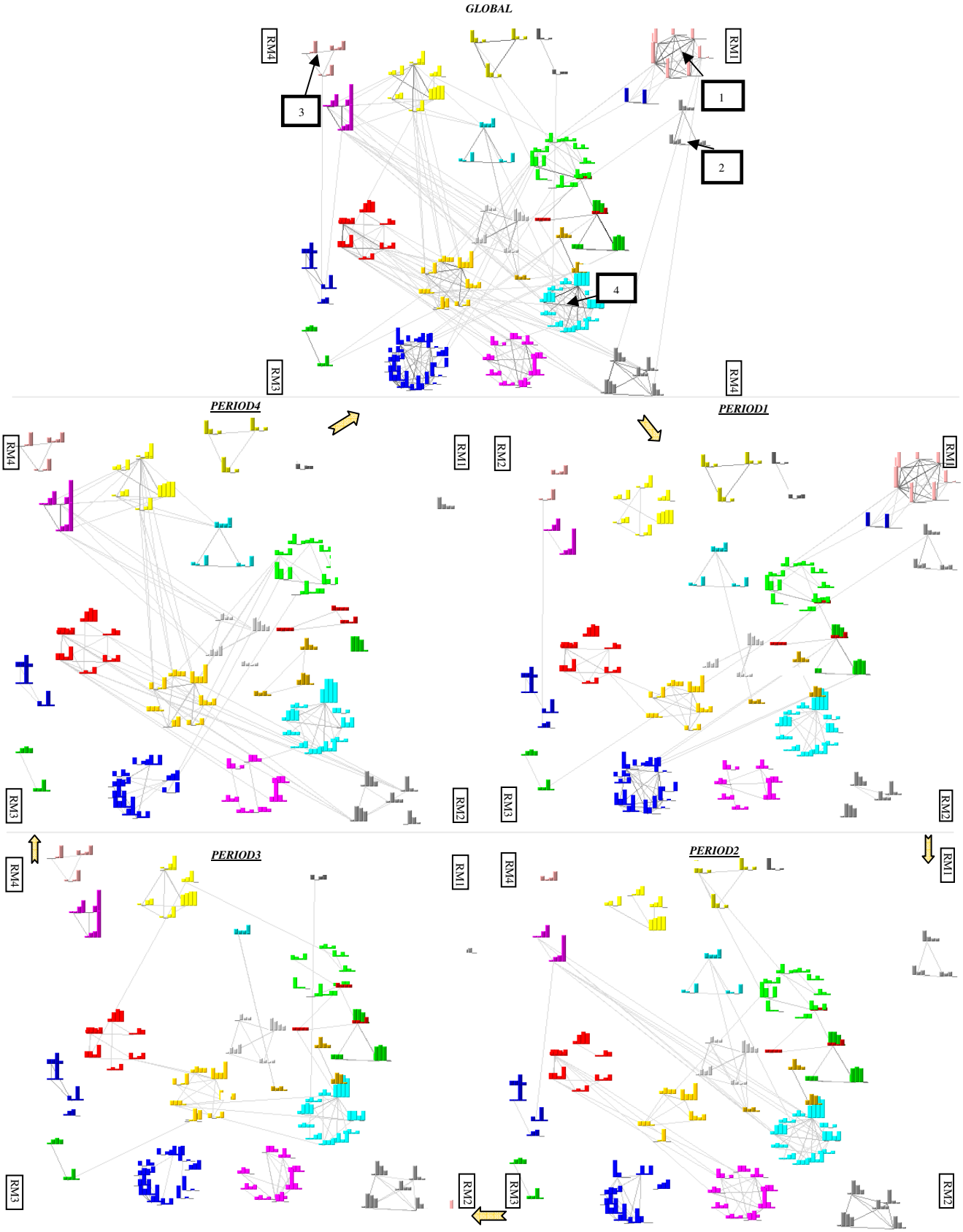

Figure 2: Temporal and reduced graphs visualizations for 4 periods.

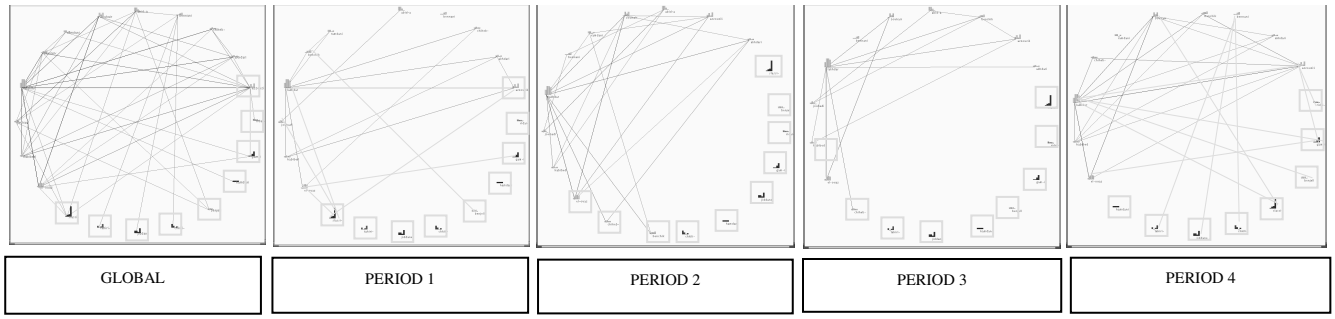

Figure 3: Extraction and evolution of the class "4" from the figure 2.

The global class "4" visualization (figure 3) is extracted from global graph (figure 2). To ease visualization, we propose a circular representation which makes it easier to analyze links between the main class ("4") and the others (in figu 3).

Whereas class "4" seems to be joined to each others classes (figure 3), if we compare the different time-sliced graphs, it is not. During the four periods, the number of class "4" components does not change but the links do. For periods 2 and 3, class "4" is not joined to the other classes. During the first and particularly the last period, class "4" is connected to components of other classes. The fourth period graph structure shows the opening to the other classes and we can suppose this class' expansion in the future.

5. CONCLUSION AND FUTURE WORK

Taking into account environmental changes is crucial for understanding human and technical issues facing a domain. Knowledge visualization in the form of large graph makes this possible. Visual depictions of graphs and networks are external representations that use human visual processing to reduce the cognitive load of many tasks required for understanding global or local structure. In this context, VisuGraph tool offers the possibility to represent data with large graphs. In this article we have presented solutions to reduce large graph complexity by adapted clustering and by taking into account the temporal dimension.

In a first step, a large graph is reduced by the use of the MCL algorithm. Only one specific node represents each class. In a second step, based on a global graph, each time-sliced graph is visualized, specifying and making data analysis easier. By this way, a non expert user is able to:

- Have qualitative and quantitative information at the appropriate time;
- Anticipate the evolution and trends of the actors on a specific domain;
- Know the competitors and determine their positioning;
- Identify newcomers on the studied domain;
- Identify collaborations, alliances…,

However the Markov Clustering does not take the user point of view into account. Partitioning could either be too fine or too coarse (in some specific case only one class may be found). It would be advisable to make the user intervene within the clustering by a visual analysis of the groupings. If attraction forces are privileged, groups can be formed, taking into account the transitivity. It would be necessary to have recourse to partitioning or clustering with traditional techniques based on nodes co-ordinates. Visual choice of the good adjustment for the nodes position in space would reveal the most significant relations, while being based on the traditional concept of Euclidean distance.

6. REFERENCES

[1]. Loubier E., Bahsoun W., Dousset B. Visualisation de l'évolution des informations relationnelles par morphing de graphe. Journées Francophones Extraction et Gestion de Connaissances Namur, Belgique, 23/01/2007-26/01/2007, *Cépaduès Editions*, pp. 43-54. (2007).

[2]. Van Dongen S. Graph Clustering by Flow Simulation. Thesis, Utrecht University, Germany. (2000).

[3]. Eades P. A heuristic for Graph Drawing. *Congressus Numerantium,* vol. 42, pp. 149-160. (1984).

[4]. Fruchterman TMJ., Reingold EM. Graph drawing by force_directed placement. *Software – Practice and experience, 21,* pp. 1129-1164. (1991),

[5]. Frick A., Ludwig A., Lehldau H. A fast adaptative layout algorithm for undirected graphs. *Proceeding of Graph Drawing 894,* pp. 388-403. (1994).

[6]. Dorogovtsev S.N., Mendes J.F.F. Effect of the accelerating growth of communications networks on their structure. *Physical Review E 63*, 025101. (2001)

[7]. Newman M.E.J. The Structure and Function of Networks. Santa Fe Institute, Santa Fe, USA. (2003).

[8]. Barabàsi A.-L., Jeong H., Neda Z., Ravasz E., Schubert A., Vicsek T. Evolution of the social network of scientific collaboration. Physica A 311 (3–4), pp. 590-614. (2002).

[9]. Dorogovtsev S. N., Mendes J. F. F. Evolution of Networks, *Advances in Physics*, n° 51, pp. 1079-1187. (2002).

[10]. Alpert C.J., Kahng A.B. Recent developments in netlist partitioning: A survey. *The VLSI journal*, vol. 19, pp.1-18. (1995).

[11]. Kuntz P., Henaux F. (2000). Numerical comparaison of two spectral decomposition for vertex clustering. Data Analysis, Classification and Related Methods*, Proceeding Of IFCS'2000, Springer Verlag, pp.581-586. (2000).*

[12]. Jouve B., Kuntz P., Velin F. Extraction de structures macroscopiques dans des grands graphes par une approche spectrale. ECA, Hermès Science publication édition, vol. 1, pp. 173-184. (2001).

[13]. Karypis G., Kumar V. (1998). Multilevel k-way partitioning scheme for irregular graphs. *Journal of Parrallel and distributed Computing,* vol. 48, pp.96-129. (1998).

[14]. Karouach S., Dousset B. Les graphes comme représentation synthétique et naturelle de l'information relationnelle de grandes tailles. Dans : *Workshop sur la recherche d'information : un nouveau passage à l'échelle, associé à INFORSID'2003*, Nancy, INFORSID, pp. 35-48. (2003).

[15]. Enright A.J., Van Dongen S. and Ouzounis C.A. An efficient algorithm for lagre-scale detection of protein families. Nucleic Acids Research, vol. 30, pp. 1575-1584. (2002).

[16]. Loubier E., Dousset B. La prise en compte de la dimension temporelle dans la visualisation de données par morphing de graphe, *VSST 2007*, à paraître. (2007).