Aggregation of Social Networks by Divisive Clustering Method

Amine Louati and Yves Lechavellier

INRIA Paris-Rocquencourt, France

{Alzennyr.Da_Silva, Yves.Lechevallier, Fabrice.Rossi}@inria.fr

Marie-Aude Aufaure

Centrale Paris, France

Marie-Aude.Aufaure@ecp.fr

HCSDA'11 Beijing October 2011

Outline

- Introduction / Motivations
- Objectives
- K-SNAP algorithm
- Our approach
- Conclusion

Introduction / motivations

The data manipulated in an enterprise context are structured data (BD) but also unstructured data such as e-mails, documents,..

Graph model is a natural way of representing and modeling structured and unstructured data in a unified manner.

The main advantage of graph model resides in its dynamic aspect and its capability to represent relations between individuals.

However, graph extracted has a huge size which makes it difficult to analyze and visualize these data also an aggregation step is needed to have more understandable graphs in order to allow user discovering underlying information and hidden relationships between clusters of individuals.

Objectives

- **Create a data model associated to a social networks**
- **Propose an aggregative approach which reduces this information.**

Descriptions of the individuals (nodes)

In social networks we have a set of individuals described by a vector of variables (numerical, categorical or symbolic) and a set of relationships.

Set of individuals $V = \{v_1, \cdots, v_n\}$

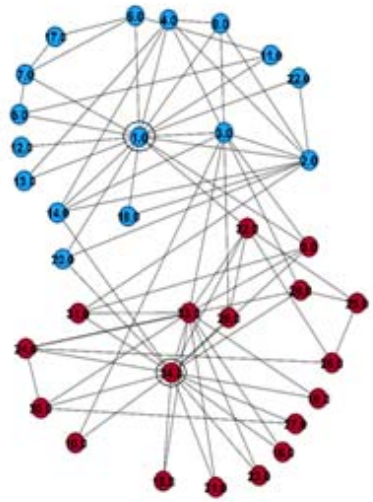
Set of relations $R = \{R_1, \cdots, R_p\}$ defined on V

Set of edges $E = \{E_1, \cdots, E_p\}$

$$u, v \in V \quad (u, v) \in E_i \text{ if } uR_iv$$

Set of variables $A = \{A_1, \cdots, A_q\}$ defined on V

Categorical variable / Relation

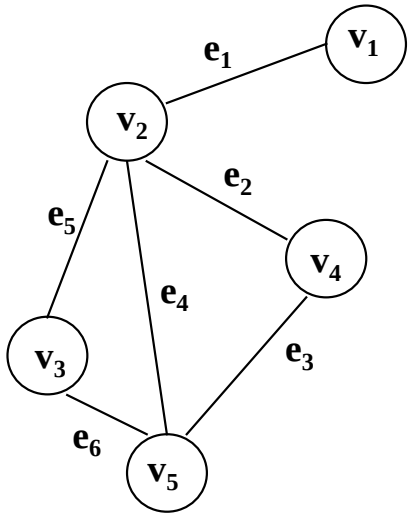


The relation “color” of individuals is a categorical variable because the relation is transitive

But the relation “call with” is not transitive also “call with” is not a categorical variable.

Zachary’s karate club dataset (UCI datasets)

Node vector space model



$$\begin{array}{c} \\ e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{array} \begin{array}{c} [v_1 \quad v_2 \quad v_3 \quad v_4 \quad v_5] \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \end{array} \quad b_2 \quad \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \quad = \quad r_2 \quad \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Build the edge-by-node matrix R

$$r_{ij} = \begin{cases} 1 & \text{if node } v_j \text{ is incident with the edge } e_i \\ 0 \end{cases}$$

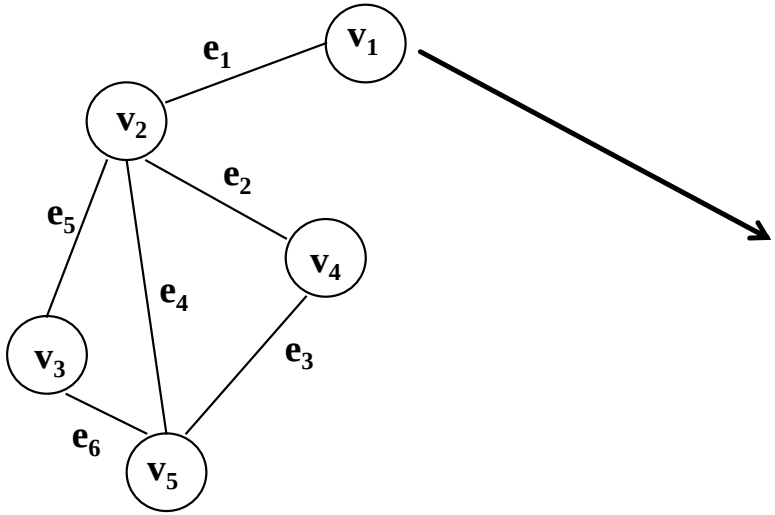
$$r_j = Rb_j$$

node vector r_j

Inria INVENTEURS DU MONDE NUMÉRIQUE

CENTRALE PARIS

Node data table model



Description data table

$$\begin{array}{c} \\ v_1 \\ \vdots \\ v_5 \end{array} \begin{bmatrix} A_1 & \cdots & A_q \\ a_1^1 & \cdots & a_1^q \\ \vdots & a_i^j & \vdots \\ a_5^1 & \cdots & a_5^q \end{bmatrix}$$

$$V = \{v_1, v_2, v_3, v_4, v_5\}$$

Each individual of V is characterized by a vector

K-means approach

The node vector v_j represents a node v_j with respect to the edges in the given graph $G=(V,E)$

The mean vector or the centroid of the node vectors contained in the cluster C_k is

$$g_k = \frac{1}{|C_k|} \sum_{v_i \in C_k} r_i$$

The objective function minimized is

$$Q_E = \sum_{k=1}^{K} \sum_{v_i \in C_k} \|r_i - g_k\|^2$$

Problems :

- The dimensional representation space is high
- How to add the data table describing the nodes ? by weight between Q_E and Q_A (objective function on description data table)?

This approach is not realist

Dissimilarity approach

The dissimilarity between two nodes is determined by the number of edges between them and a description vector of these nodes.

Let $N_{R_t}(v) = \{w \in V \mid (u,w) \in E_i\} \cup \{v\}$

the neighborhood set of the node v for the relationship R_t

For each pair (n,m) of nodes of a given the relationship R_t we compute the contingency table

	Objet m	
Objet n	a	b
	c	d

$$a = \mid N_{R_t}(m) \cap N_{R_t}(n) \mid, b = \mid N_{R_t}(m) \mid -a, c = \mid N_{R_t}(n) \mid -a,$$

Distances or dissimilarities are defined by a,b,c,d parameters

Dissimilarity approach

The most popular measures are Euclidian distance and Jaccard index which are defined by

Euclidian distance $d_1(n,m) = (b+c)/(a+b+c+d)$

	Objet m	
Objet n	a	b
	c	d

Jaccard distance

$$d_2(n,m) = 1 - a/(a+b+c)$$

Remark : with the **node vector** representation Jaccard distance is defined by:

$$d_2(n,m) = 1 - r_i^T r_j / (r_i^T r_i + r_j^T r_j - r_i^T r_j)$$

Dissimilarity clustering approach

On the set A of variables we compute dissimilarities between two nodes adapted to the different types of variables (numerical, categorical, symbolic, functional)

We have a set of data tables also we propose to use multiple dissimilarity tables clustering approach to solve this problem.

F.A.T De Carvalho, Y. Lechevallier and Filipe M. de Melo (2012). *Partitioning hard clustering algorithms based on multiple dissimilarity matrices.* Pattern Recognition.

K-SNAP algorithm

K-SNAP is a:

- Algorithm for graph aggregation based on the descriptions of nodes and edges.
- Allows the user to intervene in the aggregation procedure.

algorithm *:*

- step 1: *setting* : the user selects variables (description of the nodes), relations (description of the edges) and fix the size of the aggregated graph (number of the clusters).
- step 2: *Graph Aggregation* procedure consists of two completely independent steps:
  - Aggregation based on variable set : *A-groupement*
  - Aggregation based on relation set: *(A,R)-groupement*

Groupement concepts

A-groupement

All nodes belonging to a cluster must have the same values on all variables.

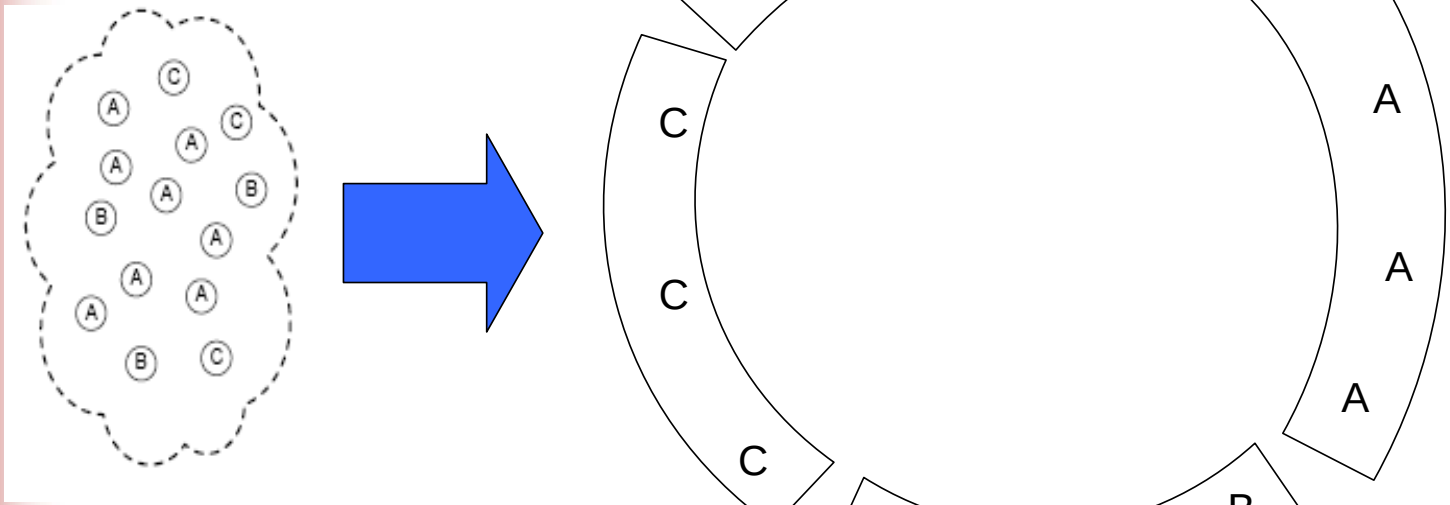
(A, R)-groupement

All nodes belonging to a cluster must have the same list of neighbor clusters.

Y. Tian, R. A. Hankins and J. M. Patel (2008). *Efficient aggregation for graph summarization***.** In SIGMOD '08

A-groupement step

3 modalities A,B,C

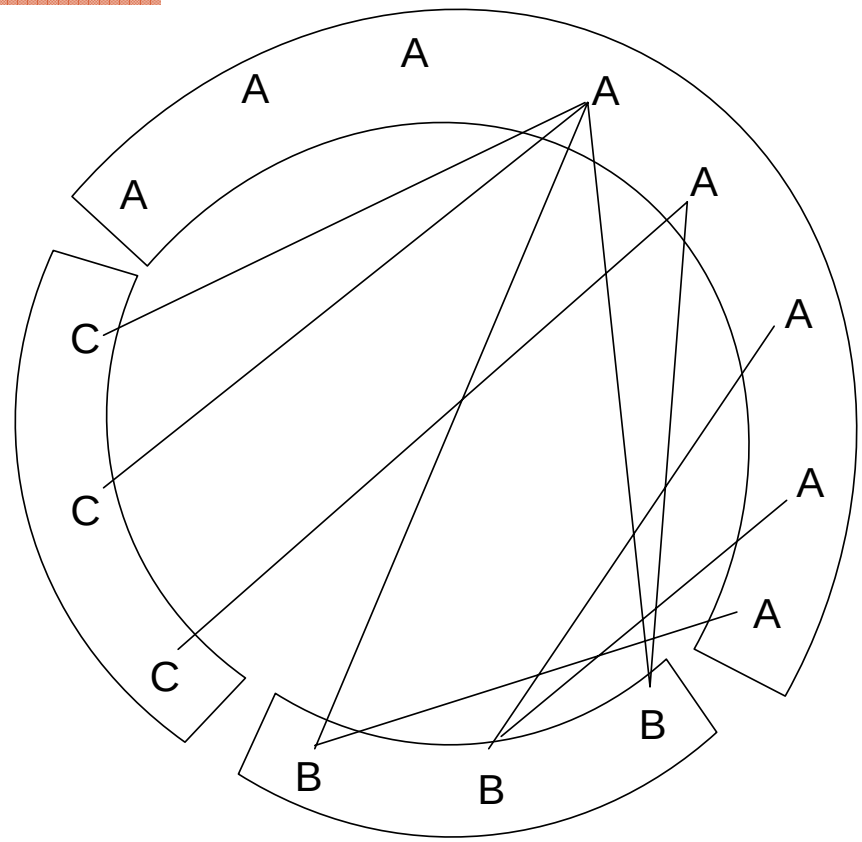


(A-R) groupement : selection step

The edge set is added.

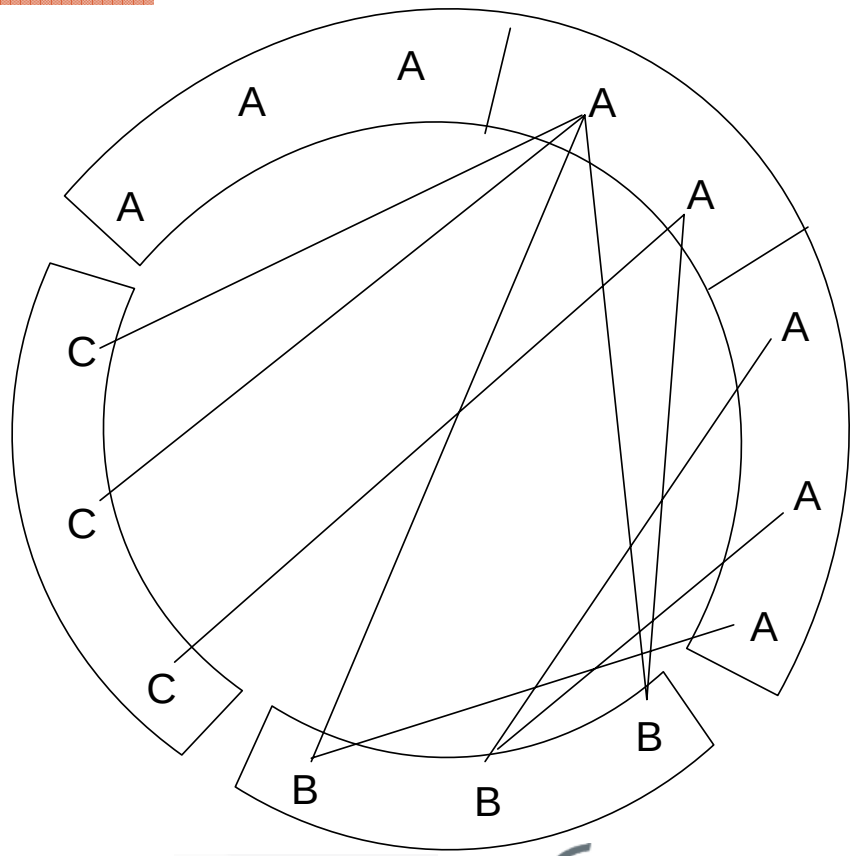
Select the cluster will must be splitting

We select the cluster A by using objective function

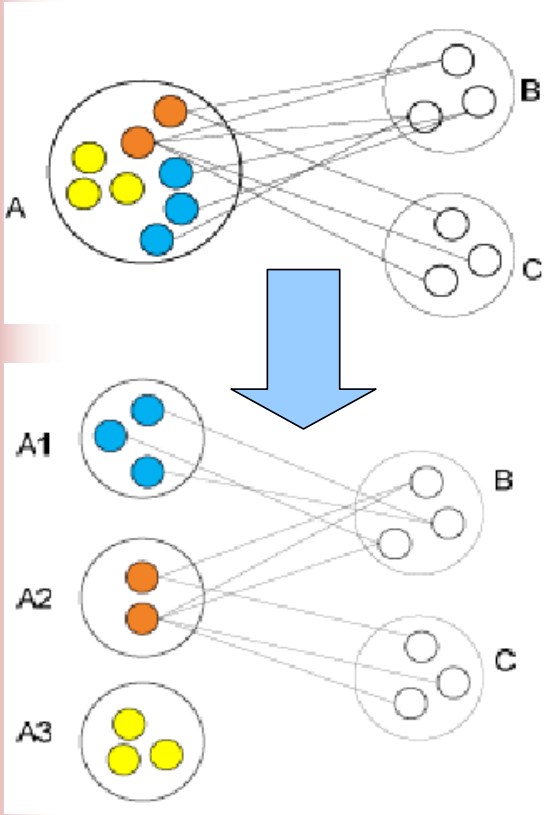


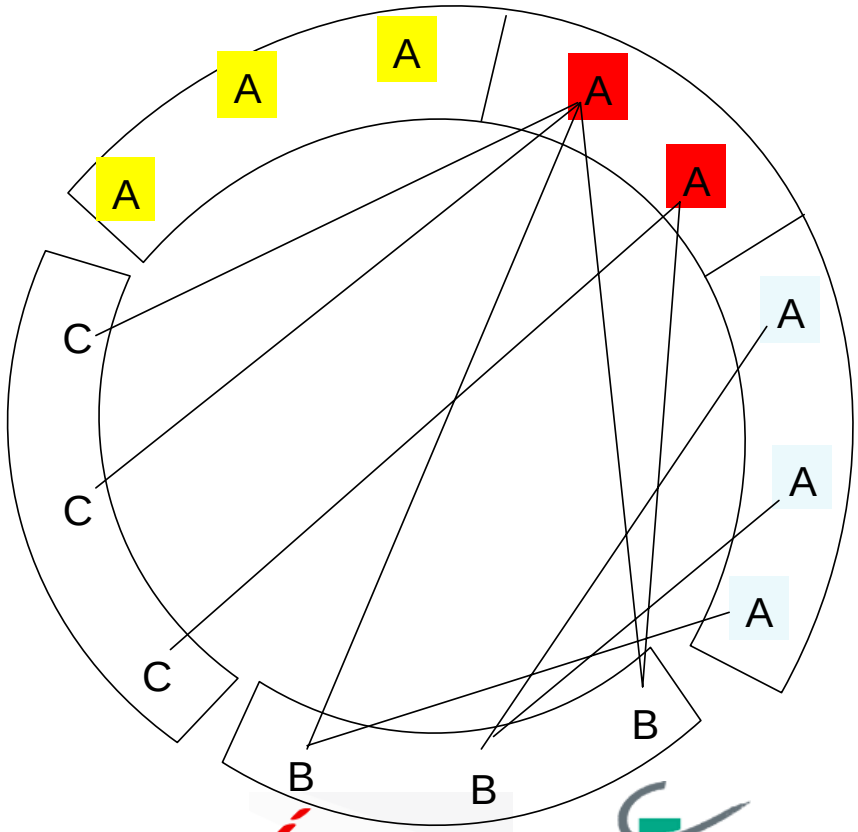
(A-R) groupement : spliting step

Divide the set A in subsets where the nodes have the same neighbor clusters



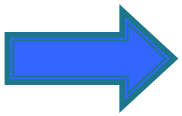
(A-R) groupement : spliting step





Limitations of k-SNAP

- Only applies to a **homogeneous** graph: nodes have the same description
- Aggregation is very rigid in terms of
  - categorical variables : **Cartesian product of all modalities**.
  - Neighbor clusters : the subsets created must be have the same neighbor clusters
- Ineffective with the presence of a large number categorical variables and heterogenic relationships.



increases the number of clusters with small size

Our approach

Integration of the clustering method "*Dynamic clustering*" in A-groupement step.

Use classical Dynamic clustering or K-means in case it has no a priori knowledge on the nodes.

Use Symbolic Dynamic clustering on the set of modalities created by A-groupement step (reduce the number of clusters)

Proposal two new aggregation criteria of evaluation to improve the quality of results while adopting the principle of k-SNAP in (A-R)-groupement step

Use the degree of node and centrality criterion

Local degree of a node

The local degree of the node v associated with the relationship R_j and the class C_i is

$$Deg_{j,i}(v) = \left| N_{R_j}(v) \cap C_i \right|$$

where $N_{R_t}(v) = \{w \in V \mid (u,w) \in E_i\} \cup \{v\}$

The complementary local degree of the node v associated with the relationship R_j and the class C_i is

$$\overline{Deg}_{j,i}(v) = \left| N_{R_j}(v) \cap \overline{C}_i \right|$$

It includes the rest of the links issued from v

Measure of homogeneity

- For a given partition $P = (C_1, C_2.., C_k)$, this measure Δ evaluates the homogeneity of the partition P and determines the cluster to be divided .
- For each relation R_j and the cluster C_i , we denote:

$$IA^j(C_i) = \frac{1}{|C_i|} \sum_{v \in C_i} Deg_{j,i}(v)$$ Intra-group criterion

$$IE^j(C_i) = \frac{1}{|E \cap C_i|} \sum_{v \in E \cap C_i} Deg_{j,i}(v)$$ Inter-group criterion

$$\Delta = \sum_{i=1}^{k} \sum_{E_j \in R} \frac{IA^j(C_i)}{IE^j(C_i)} = \sum_{i=1}^{k} \sum_{E_j \in R} \delta_i^j$$ Measure of homogeneity

with $Deg_{j,i}(v)$ is the degree of vertex v according to the relationship R_j

(A-R) groupement : selection step

R_t

The algorithm consists of finding for each iteration the relationship R and the cluster C that minimize the measure of evaluation δ until the cardinal of the partition is equal to K.

Choose the cluster i^* and the relationship j^* such that :

$$(i^*, j^*) = \arg\min_{1 \le i \le |P|, 1 \le j \le |R|} \delta_i^j = IA^j(C_i) / IE^j(C_i)$$

Measure of homogeneity

(A-R) groupement : spliting step R_t δ_i^t

On the selected cluster C_i we find the central node v_d which maximizes the centrality degree

$$d = \arg\max_{v \in C_i} Deg_{j,i}(v)$$

C_i is divided into two subgroups according to the following strategy:

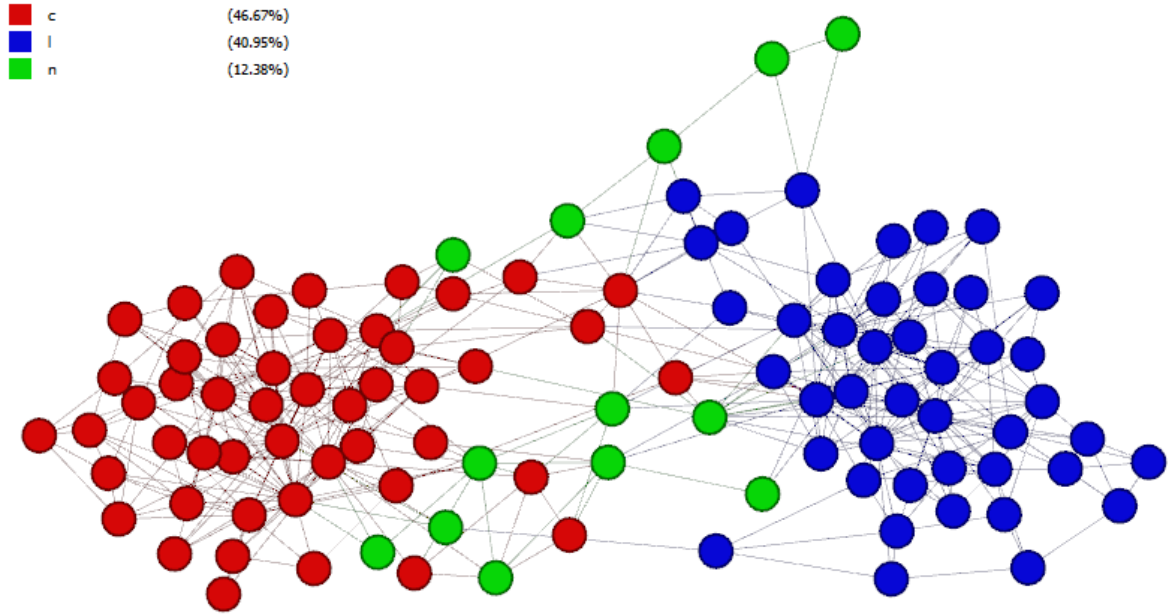
- one contains the central node with its neighbors in C_i, $N_{R_t}(v) \cap C_i$
- the other the rest of the group.

Inria INVENTEURS DU MONDE NUMÉRIQUE

CENTRALE PARIS

Application example : the network of books about US politics

Elaborated by Mark Newman this data set contains 105 vertices and 441 edges.

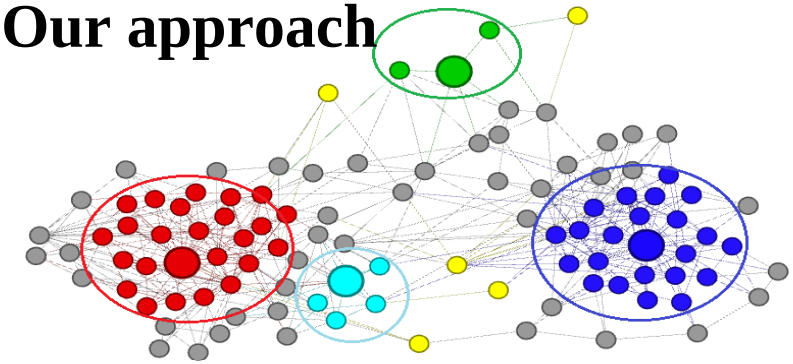


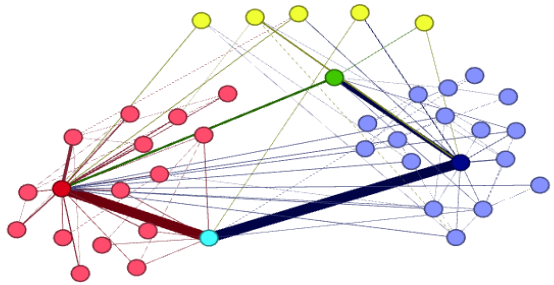
Inria INVENTEURS DU MONDE NUMÉRIQUE

CENTRALE PARIS

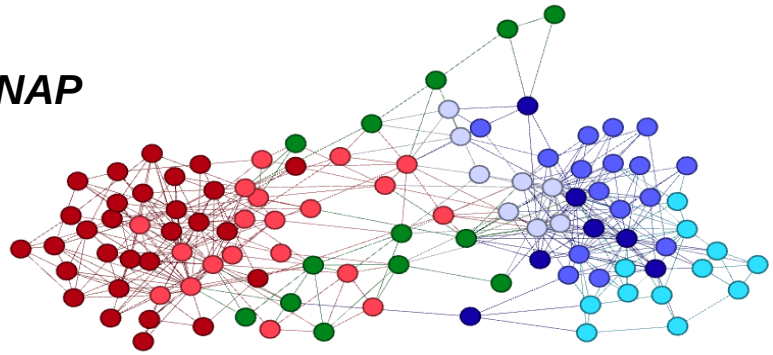
Application example : the network of books about US politics

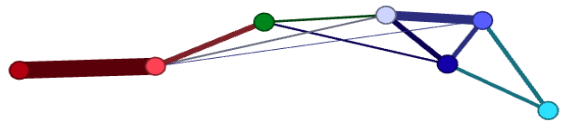
Our approach





K-SNAP





Inria INVENTEURS DU MONDE NUMÉRIQUE

CENTRALE PARIS

Conclusions

Development of new evaluation criteria to improve the quality of results by using the measure of homogeneity.

For graphs without a priori information replace the A-groupement by a clustering step

References

1 Y. Tian, R. A. Hankins and J. M. Patel (2008). *Efficient aggregation for graph summarization***.** In SIGMOD ’08

2 A Louati et al, (2011) *Recherche de classes dans les réseaux sociaux.* SFC 2011.

3 F.A.T De Carvalho, Y. Lechevallier and Filipe M. de Melo (2010). *Partitioning hard clustering algorithms based on multiple dissimilarity matrices.* Pattern Recognition.

4 R. Soussi *et al*, *Extraction et analyse de réseaux sociaux issus de bases de données relationnelles*. EGC 2011: 371-376

5 R. Godin, R. Missaoui and H. Alaoui, *Incremental concept formation algorithms based on Galois Lattices*, Computational intelligence, 11, n° 2, pp. 246 267, (1995).

Inria INVENTEURS DU MONDE NUMÉRIQUE

CENTRALE PARIS

Thank,