Mining a medieval social network by kernel SOM and related methods

Nathalie Villa-Vialaneix
http://www.nathalievilla.org
Institut de Mathématiques de Toulouse (Univ. Toulouse)
& IUT de Carcassonne (Univ. Perpignan VD)
France

Joint work with **Fabrice Rossi**, INRIA, Rocquencourt, France
and **Quoc-Dinh Truong**, IRIT, Toulouse, France

Journées MASHS, June 5*th*, 2008

Table of contents

A multidisciplinary project: "Graph-Comp"

Laboratoire d'histoire
(Univ. Le Mirail & CNRS)

Institut de Mathématiques de Toulouse
(Univ. de Toulouse & CNRS)

Institut de Recherche en Informatique
de Toulouse
(Univ. de Toulouse & CNRS)

Laboratoire d'Informatique
de Nantes Atlantique
(Univ. de Nantes & CNRS)

INRIA Rocquencourt
Projet AxIS

A multidisciplinary project: "Graph-Comp"

Florent Hautefeuille

FRAMESPA (Univ. Le Mirail)

Bertrand Jouve

IMT (Univ. Le Mirail)

Romain Boulet

IMT (Univ. Le Mirail)

Pascale Kuntz

LINA (Univ. Nantes)

Fabien Picarougne

LINA (Univ. Nantes)

Bleuenn Le Goffic

LINA (Univ. Nantes)

Taoufiq Dkaki

IRIT (Univ. Le Mirail)

Quoc-Dinh Truong

IRIT (Univ. Le Mirail)

Fabrice Rossi

INRIA Rocquencourt

A multidisciplinary project: “Graph-Comp”

Florent Hautefeuille

FRAMESPA (Univ. Le Mirail)

Bertrand Jouve

IMT (Univ. Le Mirail)

Romain Boulet

IMT (Univ. Le Mirail)

Pascale Kuntz

LINA (Univ. Nantes)

Fabien Picarougne

LINA (Univ. Nantes)

Bleuenn Le Goffic

LINA (Univ. Nantes)

Taoufiq Dkaki

IRIT (Univ. Le Mirail)

Quoc-Dinh Truong

IRIT (Univ. Le Mirail)

Fabrice Rossi

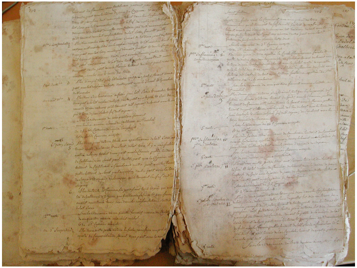
INRIA Rocquencourt

Study of a medieval corpus

Work already presented in MASHS 2007 : [Boulet et al., 2007]

A huge corpus of medieval documents

In the archives of Cahors (Lot), corpus of 5000 agrarian contracts. These contracts

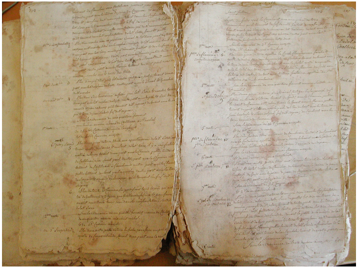


Work already presented in MASHS 2007 : [Boulet et al., 2007]

A huge corpus of medieval documents

In the archives of Cahors (Lot), corpus of 5000 agrarian contracts. These contracts

- come from four seignories (about 10 villages) of South-West of France;

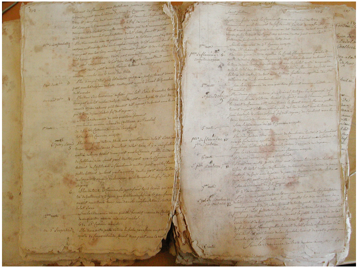


Work already presented in MASHS 2007 : [Boulet et al., 2007]

A huge corpus of medieval documents

In the archives of Cahors (Lot), corpus of 5000 agrarian contracts. These contracts

- come from four seignories (about 10 villages) of South-West of France;
- were established between 1240 and 1520 (before and after the Hundred Years' war);



Study of a medieval corpus

Work already presented in MASHS 2007 : [Boulet et al., 2007]

A huge corpus of medieval documents

In the archives of Cahors (Lot), corpus of 5000 agrarian contracts. These contracts

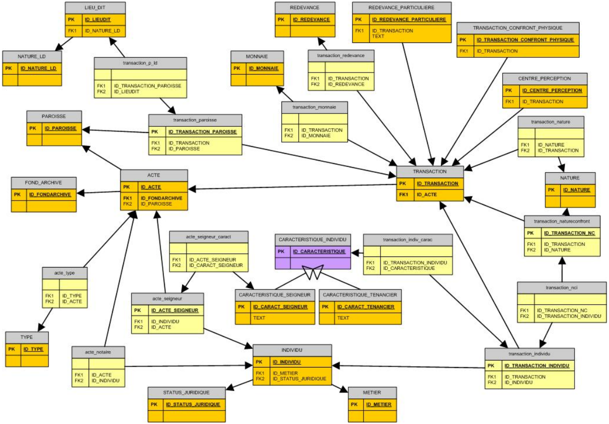
- come from four seignories (about 10 villages) of South-West of France;
- were established between 1240 and 1520 (before and after the Hundred Years' war);

This corpus interests the historians because:

- only a few documents from middle ages deal with peasants' life;
- it permits to study *a priori* the evolution of the social network before and after the Hundred Years' War.

Methodology

Each contract of the corpus is recorded in a database (still to be finished) thought by Fabien Picarougne:



Parts of this database can be accessed on the web site http://graphcomp.univ-tlse2.fr/.

A large graph for the medieval social network

From part of the database (1000 contracts before the Hundred Years' War), we built a weighted graph:

- vertices: the peasants found in the contracts (nobilities are removed);

A large graph for the medieval social network

From part of the database (1000 contracts before the Hundred Years' War), we built a weighted graph:

- vertices: the peasants found in the contracts (nobilities are removed);
- two peasants are linked together if:
  - they appear together in the same contract;
  - they appear in two different contracts which differ from less than 15 years and on which they are related to the same lord or to the same notary.

A large graph for the medieval social network

From part of the database (1000 contracts before the Hundred Years' War), we built a weighted graph:

- vertices: the peasants found in the contracts (nobilities are removed);
- two peasants are linked together if:
  - they appear together in the same contract;
  - they appear in two different contracts which differ from less than 15 years and on which they are related to the same lord or to the same notary.
- the graph thus have weights $(w_{i,j})_{i,j=1\dots,n}$ which are the number of contracts satisfying one of these conditions. They are such that:
  - $w_{i,j} = w_{j,i} \geq 0$
  - $w_{i,i} = 0$.

A large graph for the medieval social network

From part of the database (1000 contracts before the Hundred Years' War), we built a weighted graph:

- vertices: the peasants found in the contracts (nobilities are removed);
- two peasants are linked together if:
  - they appear together in the same contract;
  - they appear in two different contracts which differ from less than 15 years and on which they are related to the same lord or to the same notary.
- the graph thus have weights $(w_{i,j})_{i,j=1\ldots,n}$ which are the number of contracts satisfying one of these conditions. They are such that:
  - $w_{i,j} = w_{j,i} \geq 0$
  - $w_{i,i} = 0$.

Providing tools to help historians understanding the structure of this social network.

First description of the graph

The largest connected component of the medieval social network:

- has 615 vertices (i.e., 615 different peasants were found in the contracts),

First description of the graph

The largest connected component of the medieval social network:

- has 615 vertices (i.e., 615 different peasants were found in the contracts),
- has 4193 edges whose weights sum at 40 329 but 50% of the edges have a weight 1 and less than 2% have a weight greater than 100,

First description of the graph

The largest connected component of the medieval social network:

- has 615 vertices (i.e., 615 different peasants were found in the contracts),
- has 4193 edges whose weights sum at 40 329 but 50% of the edges have a weight 1 and less than 2% have a weight greater than 100,
- is a “small world graph” with a small density (2.2%) and a high local connectivity (77%),

First description of the graph

The largest connected component of the medieval social network:

- has 615 vertices (i.e., 615 different peasants were found in the contracts),
- has 4193 edges whose weights sum at 40 329 but 50% of the edges have a weight 1 and less than 2% have a weight greater than 100,
- is a "small world graph" with a small density (2.2%) and a high local connectivity (77%),
- is a "scale free graph": the number of vertices having a number of links greater than k is decreasing exponentially when k increases.

First description of the graph

The largest connected component of the medieval social network:

- has 615 vertices (i.e., 615 different peasants were found in the contracts),
- has 4193 edges whose weights sum at 40 329 but 50% of the edges have a weight 1 and less than 2% have a weight greater than 100,
- is a "small world graph" with a small density (2.2%) and a high local connectivity (77%),
- is a "scale free graph": the number of vertices having a number of links greater than *k* is decreasing exponentially when *k* increases.

Caracteristics similar to those found for modern social networks.

First description of the graph

The largest connected component of the medieval social network:

- has 615 vertices (i.e., 615 different peasants were found in the contracts),
- has 4193 edges whose weights sum at 40 329 but 50% of the edges have a weight 1 and less than 2% have a weight greater than 100,
- is a "small world graph" with a small density (2.2%) and a high local connectivity (77%),
- is a "scale free graph": the number of vertices having a number of links greater than k is decreasing exponentially when k increases.

Caracteristics similar to those found for modern social networks.

But ! How to visualize and/or simplify this graph to interpret it ?

Table of contents

What is graph drawing ?

Graph drawing aims at the arrangement of the vertices and edges in order to make the representation of the graph understandable and aesthetics. See Graph Visualization Software References website:

http://www.polytech.univ-nantes.fr/GVSR/

(LINA, [Pinaud et al., 2007]).

What is graph drawing ?

Graph drawing aims at the arrangement of the vertices and edges in order to make the representation of the graph understandable and aesthetics. See Graph Visualization Software References website:

`http://www.polytech.univ-nantes.fr/GVSR/`

(LINA, [Pinaud et al., 2007]).

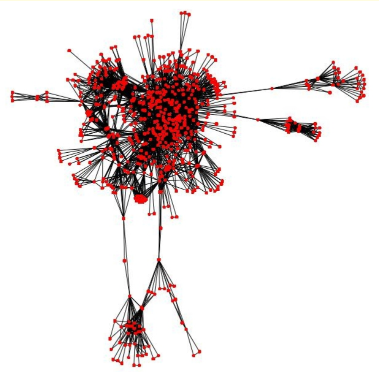
Here, Tulip.
Enables force-directed algorithms:

gradient-descent minimization of an energy function.

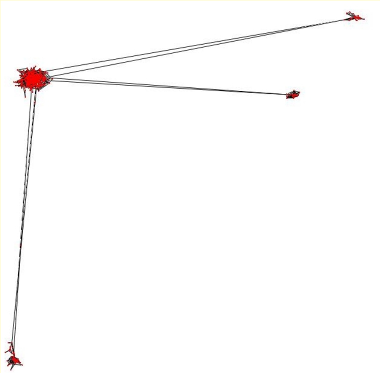
What is graph drawing ?

Graph drawing aims at the arrangement of the vertices and edges in order to make the representation of the graph understandable and aesthetics.

Representation of the medieval network by force-directed algorithms



"GEM"



"Spring Electrical"

Table of contents

Aims of the clustering

We want to underline homogeneous social groups that are fewly connected to each others

[Newman and Girvan, 2004]: *"reducing [the] level of complexity [of a network] to one that can be interpreted readily by the human eye, will be invaluable in helping us to understand the large-scale structure of these new network data"*

Aims of the clustering

We want to underline homogeneous social groups that are fewly connected to each others

[Newman and Girvan, 2004]: *“reducing [the] level of complexity [of a network] to one that can be interpreted readily by the human eye, will be invaluable in helping us to understand the large-scale structure of these new network data”*

Review on clustering of the vertices of a graph in [Schaeffer, 2007]:

- How to measure the quality of a graph clustering?
- Presentation of global or local algorithms based on
  - a similarity measure and the adaptation of a clustering algorithm to similarity data;
  - mapping of the graph on a euclidean space;
  - the minimization of a cluster fitness measures.

  Several kinds:
  - batch
  - online
  - hierarchical (divisive or agglomerative)

Aims of the clustering

We want to underline homogeneous social groups that are fewly connected to each others

[Newman and Girvan, 2004]: *"reducing [the] level of complexity [of a network] to one that can be interpreted readily by the human eye, will be invaluable in helping us to understand the large-scale structure of these new network data"*

Review on clustering of the vertices of a graph in [Schaeffer, 2007]:

- How to measure the quality of a graph clustering?
- Presentation of global or local algorithms based on
  - a similarity measure and the adaptation of a clustering algorithm to similarity data;
  - mapping of the graph on a euclidean space;
  - the minimization of a cluster fitness measures.

  Several kinds:
  - batch
  - online
  - hierarchical (divisive or agglomerative)

Spectral clustering [von Luxburg, 2007]

For a graph with vertices $V = \{x_1, \ldots, x_n\}$ having positive weights $(w_{i,j})_{i,j=1,\ldots,n}$ **Laplacian**: $L = (L_{i,j})_{i,j=1,\ldots,n}$ where

$$L_{i,j} = \begin{cases} -w_{i,j} & \text{if } i \neq j \\ d_i & \text{if } i = j \end{cases} ;$$

Spectral clustering [von Luxburg, 2007]

For a graph with vertices $V = \{x_1, \ldots, x_n\}$ having positive weights $(w_{i,j})_{i,j=1,\ldots,n}$ **Laplacian**: $L = (L_{i,j})_{i,j=1,\ldots,n}$ where

$$L_{i,j} = \begin{cases} -w_{i,j} & \text{if } i \neq j \\ d_i & \text{if } i = j \end{cases} ;$$

Graph cut optimization

If the graph is connected, clustering the vertices into k groups $A_1, \ldots, A_k$ that minimize

$$Cut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \sum_{j \in A_i,\ j' \notin A_i} w_{i,i'}$$

Spectral clustering [von Luxburg, 2007]

For a graph with vertices $V = \{x_1, \ldots, x_n\}$ having positive weights $(w_{i,j})_{i,j=1,\ldots,n}$ **Laplacian**: $L = (L_{i,j})_{i,j=1,\ldots,n}$ where

$$L_{i,j} = \begin{cases} -w_{i,j} & \text{if } i \neq j \\ d_i & \text{if } i = j \end{cases} ;$$

Graph cut optimization

If the graph is connected, clustering the vertices into k groups $A_1, \ldots, A_k$ that minimize

$$Cut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \sum_{j \in A_i,\ j' \notin A_i} w_{i,i'}$$

$\Leftrightarrow$ find $(h_1, \ldots, h_k) \in \prod_{i=1}^{k} \left\{0, \frac{1}{\sqrt{|A_i|}}\right\}^n$ that minimizes

$$\sum_{i=1}^{k} h_i^T L h_i \text{ subject to } (h_1 \ldots h_k)(h_1 \ldots h_k)^T = \mathbb{I}_n$$

Spectral clustering [von Luxburg, 2007]

For a graph with vertices $V = \{x_1, \ldots, x_n\}$ having positive weights $(w_{i,j})_{i,j=1,\ldots,n}$ **Laplacian**: $L = (L_{i,j})_{i,j=1,\ldots,n}$ where

$$L_{i,j} = \begin{cases} -w_{i,j} & \text{if } i \neq j \\ d_i & \text{if } i = j \end{cases};$$

Graph cut optimization

If the graph is connected, clustering the vertices into k groups $A_1, \ldots, A_k$ that minimize

$$Cut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \sum_{j \in A_i,\ j' \notin A_i} w_{i,i'}$$

$\simeq$ find $h_1, \ldots, h_k \in \mathbb{R}^n$ that minimize (continuous approximation)

$$\sum_{i=1}^{k} h_i^T L h_i \text{ subject to } (h_1 \ldots h_k)(h_1 \ldots h_k)^T = \mathbb{I}_n$$

Spectral clustering on medieval social network

Algorithm

1. Compute the eigenvectors, $v_1, \ldots, v_k \in \mathbb{R}^n$ of L associated with the k smallest positive eigenvalues.

Spectral clustering on medieval social network

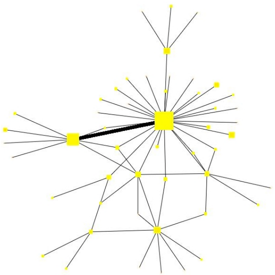
Algorithm

1. Compute the eigenvectors, $v_1, \ldots, v_k \in \mathbb{R}^n$ of L associated with the k smallest positive eigenvalues.
2. Use the rows of the matrix $(v_1 \ldots v_k)$ has a mapping of the graph in $\mathbb{R}^k$ and perform a clustering algorithm on them.

Algorithm

1. Compute the eigenvectors, $v_1, \dots, v_k \in \mathbb{R}^n$ of L associated with the k smallest positive eigenvalues.
2. Use the rows of the matrix $(v_1 \dots v_k)$ has a mapping of the graph in $\mathbb{R}^k$ and perform a clustering algorithm on them.

Representation of the clustering (k-means, 50 clusters + force directed algorithm):



2 big clusters of central people highly connected;

Identification of individuals that help to connect the network;

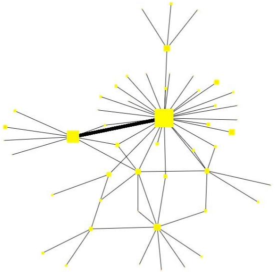
isolated individuals around.

Spectral clustering on medieval social network

Algorithm

1. Compute the eigenvectors, $v_1, \ldots, v_k \in \mathbb{R}^n$ of L associated with the k smallest positive eigenvalues.
2. Use the rows of the matrix $(v_1 \ldots v_k)$ has a mapping of the graph in $\mathbb{R}^k$ and perform a clustering algorithm on them.

Representation of the clustering (k-means, 50 clusters + force directed algorithm):



2 big clusters of central people highly connected;

Identification of individuals that help to connect the network;

isolated individuals around.

But: Size of the biggest cluster: 268 !!

16 clusters have size 1

more than 50% of the clusters have a size less than 2

A regularized version of the Laplacian: the heat kernel

From the diffusion matrix to the heat kernel

Diffusion matrix: for $\beta > 0$, $K^\beta = e^{-\beta L}$.

$\Rightarrow$

$$\begin{array}{rccc} k^\beta : & V \times V & \rightarrow & \mathbb{R} \\ & (x_i, x_j) & \rightarrow & K^\beta_{i,j} \end{array}$$

is the **diffusion kernel** (or heat kernel).

A regularized version of the Laplacian: the heat kernel

From the diffusion matrix to the heat kernel

Diffusion matrix: for $\beta > 0$, $K^\beta = e^{-\beta L}$.

$\Rightarrow$

$$\begin{array}{rrcl} k^\beta : & V \times V & \rightarrow & \mathbb{R} \\ & (x_i, x_j) & \rightarrow & K^\beta_{i,j} \end{array}$$

is the **diffusion kernel** (or heat kernel).

Intuitive interpretation: $k^\beta(x_i, x_j) \simeq$ quantity of energy accumulated in x_j after a given time if energy is injected in x_i at time 0 and if diffusion is done along the edges (β control the intensity of the diffusion).

A regularized version of the Laplacian: the heat kernel

From the diffusion matrix to the heat kernel

Diffusion matrix: for $\beta > 0$, $K^\beta = e^{-\beta L}$.

$\Rightarrow$

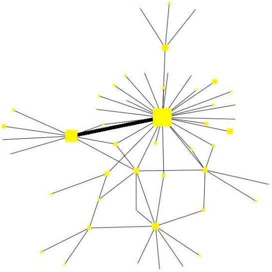
$$\begin{array}{rccc} k^\beta: & V \times V & \rightarrow & \mathbb{R} \\ & (x_i, x_j) & \rightarrow & K^\beta_{i,j} \end{array}$$

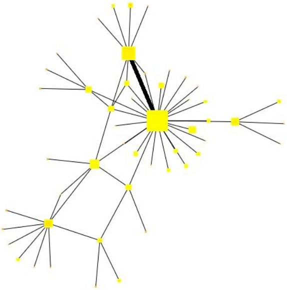
is the **diffusion kernel** (or heat kernel).

Mapping of the graph on a euclidean space: k^β is the scalar product associated with the mapping

$$\phi : x_i \in V \rightarrow (v_1 \ldots v_n)_i \in \mathbb{R}^n_\lambda$$

where $(v_l)_l$ are the eigenvectors of L and $\mathbb{R}^n_\lambda$ denotes the n-dimensional space with norm weighted by $(e^{-\beta\lambda_l})_l$ ($\lambda \equiv$ eigenvalues of L).

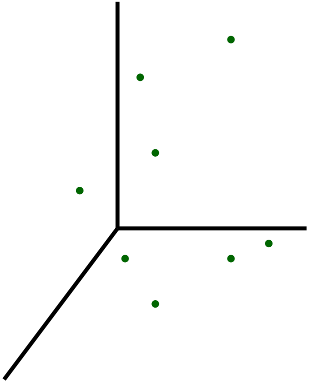




Spectral Clustering	Kernel *k*-means
Max size: 268	242
Nb of clusters of size 1: 16	17
Median size: 2	2

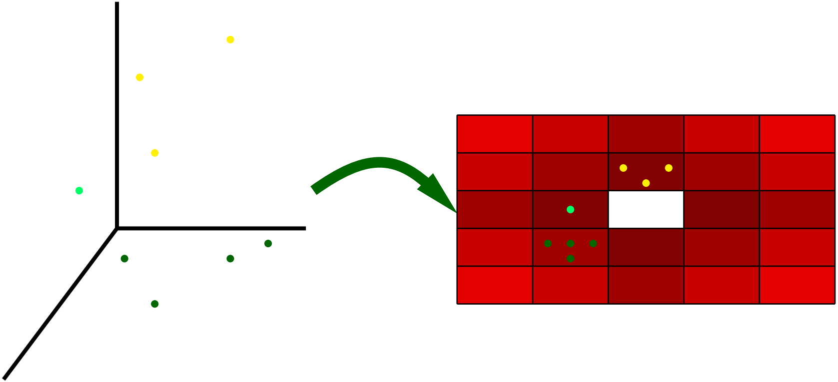
Table of contents

General principle of SOM for graphs



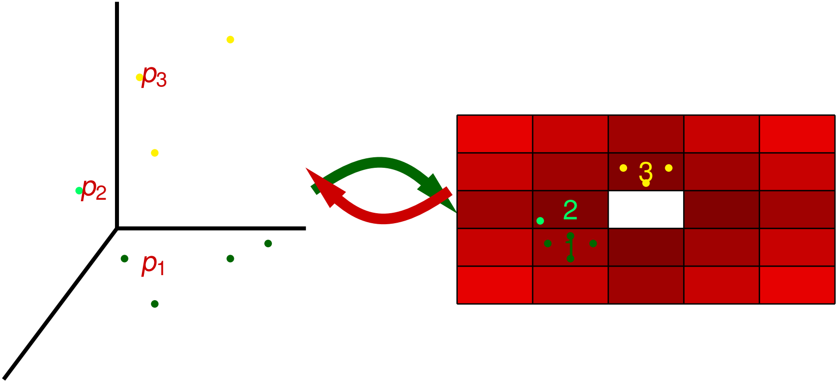
The vertices of the graph are mapped on a euclidean space (by L: "spectral SOM" or by K: "kernel SOM").

General principle of SOM for graphs



Each vertex x_i is affected to a neuron (a cluster) of the Kohonen map, $f(x_i)$. Neurons are related to each others by a neighborhood relationship ("distance": d).

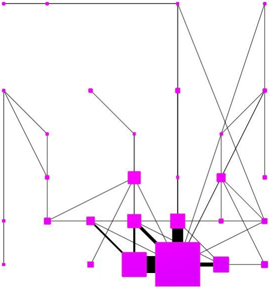
General principle of SOM for graphs

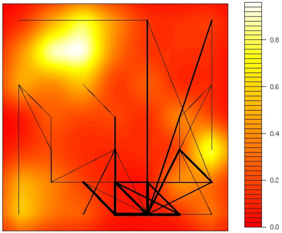


Each neuron j of the map is represented by a prototype p_j.
Couples (j, p_j) and $(x_i, f(x_i))$ depend from each others and are iteratively updated in order to approach the minimization of the energy of the map:

$$\mathcal{E}^n = \sum_{j=1}^{n} \sum_{i=1}^{M} h(d(f(x_j), i)) \|\phi(x_i) - p_j\|^2.$$

Spectral SOM



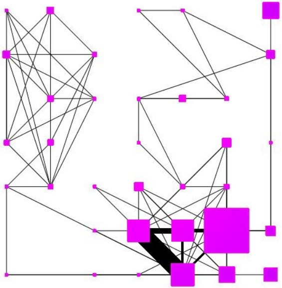


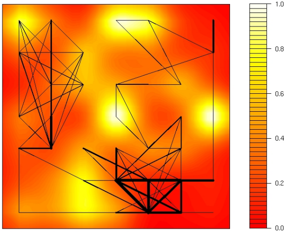
Number of clusters 29
Number of clusters of size 1 11
Maximum size of the clusters 325
Median size 2

Q-modularity: $\sum_{i=1}^{k}(e_i - a_i^2)$

$Q_{\text{modul}} = 0.433$
(vs 0.420 & 0.425 for clusterings)

Kernel SOM [Boulet et al., 2008]



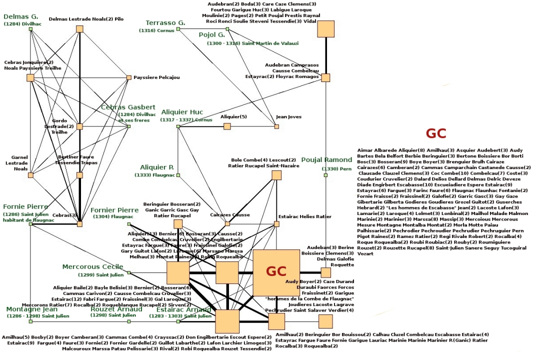


Number of clusters 35
Number of clusters of size 1 13
Maximum size of the clusters 255
Median size 3

Q-modularity: $\sum_{i=1}^{k}(e_i - a_i^2)$

$$Q_{\text{modul}} = 0.551$$

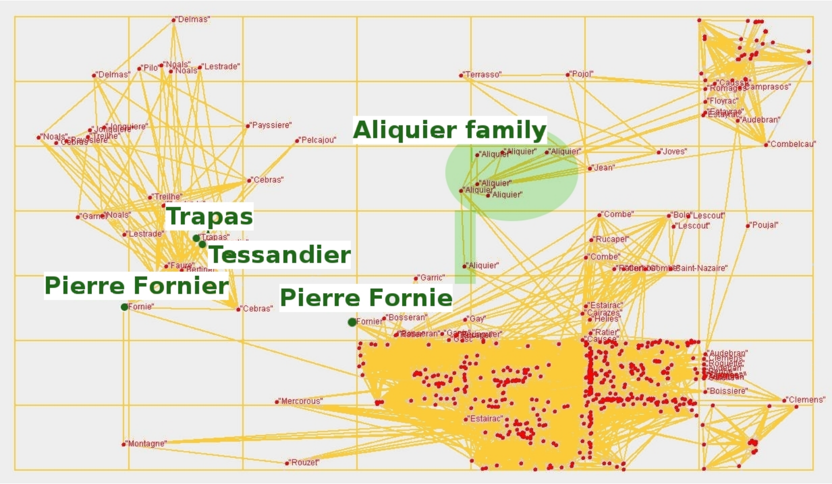
1. Dates
2. Lieux

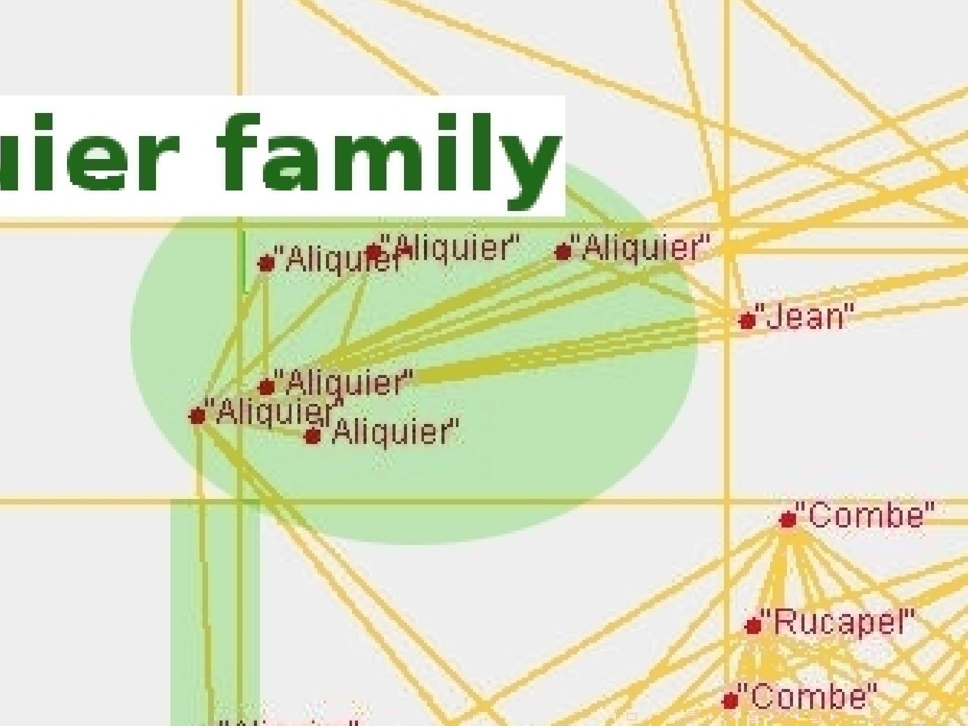


Force directed algorithm for clustered graphs [Truong et al., 2007, Truong et al., 2008]

La Suite...

By adding constrains on force-directed algorithms




uier family
"Aliquier"
"Aliquier"
"Aliquier"
"Jean"
"Aliquier"
"Aliquier"
"Aliquier"
"Combe"
"Rucapel"
"Combe"


"Ratier"
"Causse"
"Estairac"

Perspectives

Several perspectives:

- Improving the global representation of the network (hierarchical algorithms, improving algorithms for clustered graphs drawing, others algorithms such as simulated annealing, . . .)

Several perspectives:

- Improving the global representation of the network (hierarchical algorithms, improving algorithms for clustered graphs drawing, others algorithms such as simulated annealing, . . .)
- Understanding the evolution of the social network through time (before/during/after Hundred Years' War): specific tools have to be built in order to
  - understand what become the dominant families ("Aliquier", "Fornie", . . .),
  - make a comparison despite the fact that the vertices are not the same.

References

Boulet, R., Hautefeuille, F., Jouve, B., Kuntz, P., Le Goffic, B., Picarougne, F., and Villa, N. (2007).
Sur l'analyse de réseaux de sociabilité dans la société paysanne médiévale.
In *MASHS 2007*, Brest, France.

Boulet, R., Jouve, B., Rossi, F., and Villa, N. (2008).
Batch kernel SOM and related laplacian methods for social network analysis.
Neurocomputing, 71(7-9):1257–1273.

Newman, M. and Girvan, M. (2004).
Finding and evaluating community structure in networks.
Physical Review, E, 69:026113.

Pinaud, B., Kuntz, P., and Picarougne, F. (2007).
The website for graph visualization software references (GVSR).
In *Graph Drawing: 14th International Symposium on Graph Drawing 2006*, volume 4372 of *Lecture Notes in Computer Science*, pages 440–441, Berlin, Heidelberg. Springer.

Schaeffer, S. (2007).
Graph clustering.
Computer Science Review, 1(1):27–64.

Truong, Q., Dkaki, T., and Charrel, P. (2007).
An energy model for the drawing of clustered graphs.
In *Proceedings of Vème colloque international VSST*, Marrakech, Maroc.

Truong, Q., Dkaki, T., and Charrel, P. (2008).
Clustered graphs drawing.
In *Proceedings of Stimulating Manufacturing Excellence in SME*, Hammamet, Tunisie.

von Luxburg, U. (2007).
A tutorial on spectral clustering.
Statistics and Computing, 17(4):395–416.