Self-Organizing Maps for clustering and visualization of bipartite graphs

Nathalie Villa-Vialaneix

en collaboration avec Madalina Olteanu

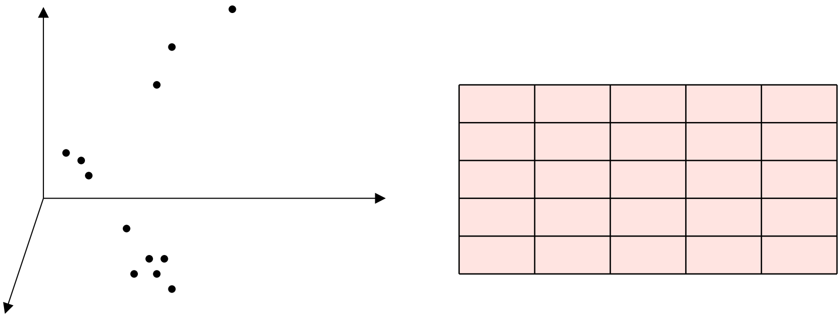
nathalie.villa@toulouse.inra.fr
http://www.nathalievilla.org

Journées de Statistique de la SFdS
Rennes, 5 juin 2014

Outline

Outline

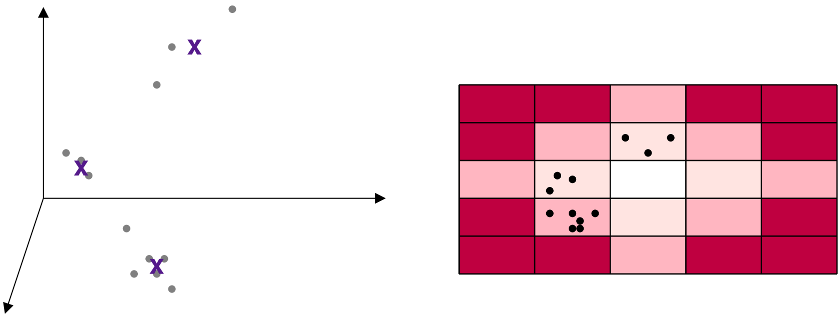
Basics on SOM [Kohonen, 2001]



Aim: Project the data $x \in \mathbb{R}^d$ on a square 2-dimensional grid made of U units $\{1, \ldots, U\}$:

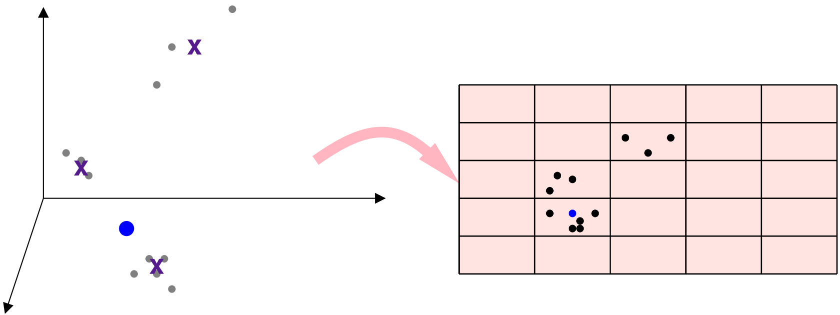
- clustering
- non-linear projection (& visualization) that preserves topology
- generalizes k-means

Basics on SOM [Kohonen, 2001]



- $(x_i)_{i=1,\ldots,n} \subset \mathbb{R}^d$ are affected to a unit $\mathcal{C}(x_i) \in \{1, \ldots, U\}$
- the grid is equipped with a "distance" between units: $d(u, u')$ and observations affected to close units are close in $\mathbb{R}^d$
- every unit u corresponds to a prototype, p_u (**x**) in $\mathbb{R}^d$

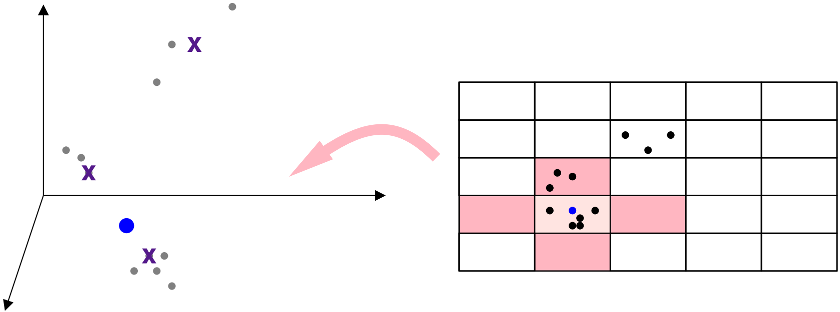
Basics on SOM [Kohonen, 2001]



Iterative learning (affectation step): x_i is picked at random within $(x_k)_k$ and affected to *best matching unit*:

$$C(x_i) = \arg\min_u \|x_i - p_u\|^2$$

Basics on SOM [Kohonen, 2001]



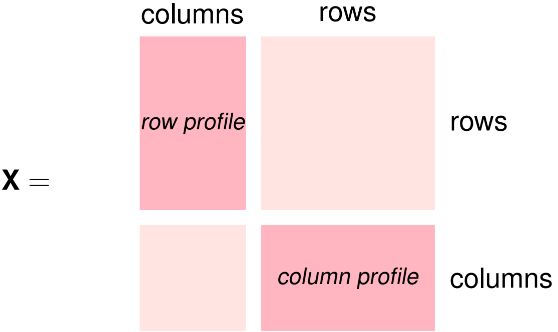
Iterative learning (representation step): all prototypes in neighboring units are updated with a gradient descent like step minimizing $\mathcal{E} = \sum_{i=1}^{n} \sum_{u=1}^{U} H^t(d(\mathcal{C}(x_i), u)) \|x_i - p_u\|^2$:

$$p_u^{t+1} \longleftarrow p_u^t + \mu(t) H^t(d(\mathcal{C}(x_i), u))(x_i - p_u^t)$$

Extensions to non vectorial data 1

KORRESP [Cottrell et al., 1993]

Data: contingency table $\mathbf{T} = (n_{ij})_{ij}$ with p rows and q columns transformed into a numeric dataset $\mathbf{X}$:



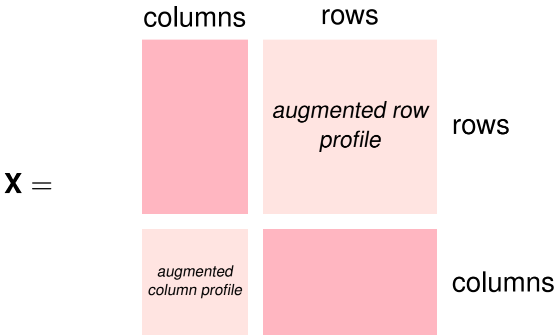
with

- $\forall\, i = 1, \ldots, p$ and $\forall\, j = 1, \ldots, q$, $\mathbf{x}_{ij} = \frac{n_{ij}}{n_{i.}} \times \sqrt{\frac{n}{n_{.j}}}$

Extensions to non vectorial data 1

KORRESP [Cottrell et al., 1993]

Data: contingency table $\mathbf{T} = (n_{ij})_{ij}$ with p rows and q columns transformed into a numeric dataset $\mathbf{X}$:



with

- $\forall\, i = 1, \ldots, p$ and $\forall\, j = q+1, \ldots, q+p$, $\mathbf{x}_{ij} = \mathbf{x}_{k(i)+p, j+q}$ with $k(i) = \arg\max_{k=1,\ldots,q} \mathbf{x}_{ik}$

Extensions to non vectorial data 1

KORRESP [Cottrell et al., 1993]

Data: contingency table $\mathbf{T} = (n_{ij})_{ij}$ with p rows and q columns transformed into a numeric dataset $\mathbf{X}$:

$\mathbf{X} =$

	columns	rows
rows	*row profile*	*augmented row profile*
columns	*augmented column profile*	*column profile*

- affectation uses reduced profile
- representation uses augmented profile
- alternatively process row profiles and column profiles

This method is implemented in the R package **SOMbrero**.

Extensions to non vectorial data 2

Relational SOM [Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ non necessarily vectorial)

Extensions to non vectorial data 2

Relational SOM [Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ non necessarily vectorial)

Adaptations of the SOM algorithm:

- prototypes: expressed as (symbolic) convex combination of $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ and $\sum_i \gamma_{ui} = 1$

Extensions to non vectorial data 2

Relational SOM [Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ non necessarily vectorial)

Adaptations of the SOM algorithm:

- prototypes: expressed as (symbolic) convex combination of $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ and $\sum_i \gamma_{ui} = 1$
- distance computation: $\|x_i - p_u\|^2$ replaced by

$$(\mathbf{D}\gamma_u)_i - \frac{1}{2}\gamma_u^T \mathbf{D}\gamma_u$$

in reference to a pseudo-Euclidean framework [Goldfarb, 1984]

Extensions to non vectorial data 2

Relational SOM [Olteanu and Villa-Vialaneix, 2014]

Data: described by a dissimilarity matrix $\mathbf{D} = (\delta(x_i, x_j))_{i,j=1,\ldots,n}$ ($(x_i)_i$ non necessarily vectorial)

Adaptations of the SOM algorithm:

- prototypes: expressed as (symbolic) convex combination of $(x_i)_i$: $p_u \sim \sum_{i=1}^n \gamma_{ui} x_i$, $\gamma_{ui} \geq 0$ and $\sum_i \gamma_{ui} = 1$
- distance computation: $\|x_i - p_u\|^2$ replaced by

$$(\mathbf{D}\gamma_u)_i - \frac{1}{2}\gamma_u^T \mathbf{D}\gamma_u$$

  in reference to a pseudo-Euclidean framework [Goldfarb, 1984]
- representation: replaced by an update of $(\gamma_u)_u$:

$$\gamma_u^{t+1} \leftarrow \gamma_u^t + \mu(t) H^t(d(\mathcal{C}(x_i), u)) \left(\mathbf{1}_i - \gamma_u^t\right)$$

  with $\mathbf{1}_{il} = 1$ if $l = i$ and 0 otherwise.

Outline

1. a short review of Self-Organizing Maps for non vectorial data
2. SOM for bipartite graphs
3. Application

Bipartite graphs

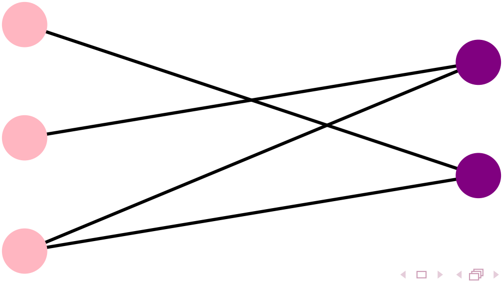
a bipartite graph is a graph $\mathcal{G} = (V, E, W, C)$ such that:

- the vertices $V = \{x_1, \ldots, x_n\}$ are split into two sets (labeled with $C_i = 0$ or $C_i = 1$);

Bipartite graphs

a bipartite graph is a graph $\mathcal{G} = (V, E, W, C)$ such that:

- the vertices $V = \{x_1, \ldots, x_n\}$ are split into two sets (labeled with $C_i = 0$ or $C_i = 1$);
- the edges $E \subset V \times V$ are such that if $(x_i, x_j) \in E \Rightarrow$ either $(C_i = 1$ and $C_j = 0)$ or $(C_i = 0$ and $C_j = 1)$;
- the edges are eventually weighted by W_{ij} (st: $W_{ii} = 0$, $W_{ij} = W_{ji} \geq 0$).



Bipartite graphs

a bipartite graph is a graph $\mathcal{G} = (V, E, W, C)$ such that:

- the vertices $V = \{x_1, \ldots, x_n\}$ are split into two sets (labeled with $C_i = 0$ or $C_i = 1$);
- the edges $E \subset V \times V$ are such that if $(x_i, x_j) \in E \Rightarrow$ either $(C_i = 1$ and $C_j = 0)$ or $(C_i = 0$ and $C_j = 1)$;
- the edges are eventually weighted by W_{ij} (st: $W_{ii} = 0$, $W_{ij} = W_{ji} \geq 0$).

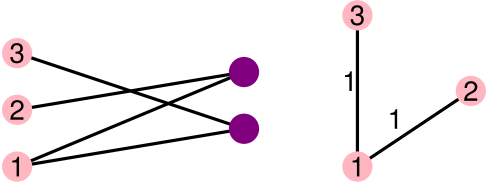
Examples of such data:

- very frequently used in recommandation systems (persons liking pages in facebook, persons buying objects...)
- authorship networks (persons and articles)
- affiliation networks (persons and firms)
- ...

Clustering bipartite graphs

Most frequent approaches are based on projected graphs, $\mathcal{G}^0$ and $\mathcal{G}^1$ st:

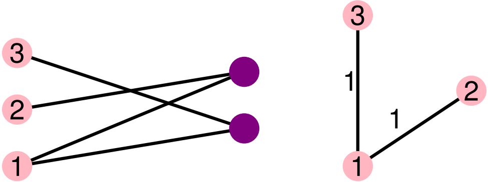
- $V^0 = \{x_i \in \mathcal{G} : \ C_i = 0\}$
- $(x_i, x_j) \in E^0 \Leftrightarrow \begin{cases} x_i, \ x_j \in V^0 \\ \exists x_k \notin V^0 : \ \{(x_i, x_k) \in E \text{ and } (x_k, x_j) \in E\} \end{cases}$
- $W^0_{ij} = \{x_k \notin V^0 : \ (x_i, x_k) \in E \text{ and } (x_k, x_j) \in E\}$



Clustering bipartite graphs

Most frequent approaches are based on projected graphs, $\mathcal{G}^0$ and $\mathcal{G}^1$ st:

- $V^0 = \{x_i \in \mathcal{G} : \ C_i = 0\}$
- $(x_i, x_j) \in E^0 \Leftrightarrow \begin{cases} x_i, \ x_j \in V^0 \\ \exists x_k \notin V^0 : \ \left\{(x_i, x_k) \in E \text{ and } (x_k, x_j) \in E\right\} \end{cases}$
- $W^0_{ij} = \left\{x_k \notin V^0 : \ (x_i, x_k) \in E \text{ and } (x_k, x_j) \in E\right\}$



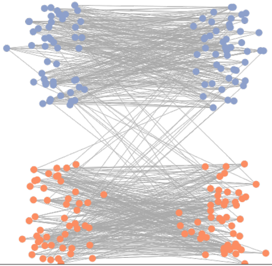
But not useful to understand the relations between the two types of nodes...

Similarities in the projected graph

simulated bipartite graph:

- nodes (labeled either 0 or 1) belong to 2 (densely connected) groups;
- edges are generated independently with a given probability: nodes within the same groups have a high probability to be connected and nodes between two groups have a low probability to be connected.

Bipartite graph:

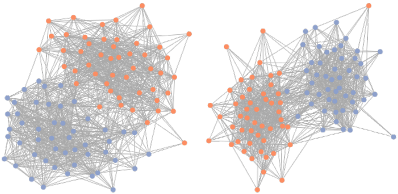


Similarities in the projected graph

simulated bipartite graph:

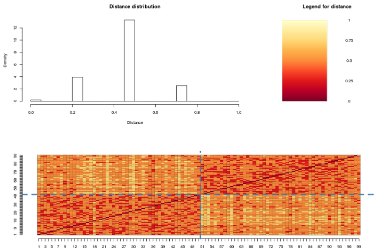
- nodes (labeled either 0 or 1) belong to 2 (densely connected) groups;
- edges are generated independently with a given probability: nodes within the same groups have a high probability to be connected and nodes between two groups have a low probability to be connected.

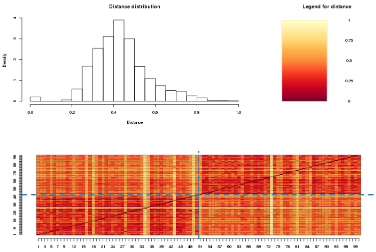
Projected graphs:



Similarities in the projected graph

simulated bipartite graph:





length of shortest paths

(based on) number of common neighb

SOM for bipartite graphs

bipartite graphs are:

- dissimilarity data;
- and data describing relations between two sets of objects (like contingency tables)

SOM for bipartite graphs

bipartite graphs are:

- dissimilarity data;
- and data describing relations between two sets of objects (like contingency tables)

⇒ adapt KORRESP: alternatively process nodes of type 0 and 1

1. pick a node at random and affect it to the closest prototype (dissimilarity SOM based on a bipartite dissimilarity)

SOM for bipartite graphs

bipartite graphs are:

- dissimilarity data;
- and data describing relations between two sets of objects (like contingency tables)

$\Rightarrow$ adapt KORRESP: alternatively process nodes of type 0 and 1

1. pick a node at random and affect it to the closest prototype (dissimilarity SOM based on a bipartite dissimilarity)
2. $p_u = (\underbrace{\sum_{i:\ c_i=0} \gamma_{ui}^0 x_i}_{\text{"type" 0 part}}, \underbrace{\sum_{i:\ c_i=1} \gamma_{ui}^1 x_i}_{\text{"type" 1 part}})$ is updated in two steps:
   - *when processing a node of type 0* standard dissimilarity update for $(\gamma_{ui}^0)_{i:\ c_i=0}$
   - *when processing a node of type 0*

$$\gamma_u^{1,t+1} \leftarrow \frac{\gamma_u^{1,t} + \mu(t)H^t(d(\mathcal{C}(x_i),u))\left(\sum_{k\in\mathcal{N}(x_i)} \mathbf{1}_k - \gamma_u^{1,t}\right)}{1+\mu(t)H^t(d(\mathcal{C}(x_i),u))(d_i^0-1)}$$

Outline

INRA SCIENCE & IMPACT

Simulated bipartite graph

100 randomly generated graphs clustered into 3 groups

- nodes (labeled either 0 or 1) belong to 3 (densely connected) groups with $\sim$ 50 nodes each;
- edges are generated independently with a given probability (high intra-group probability and low inter-group probability)

Simulated bipartite graph

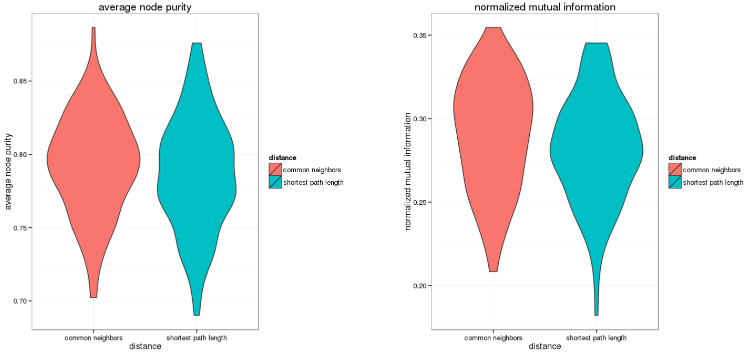
100 randomly generated graphs clustered into 3 groups

- nodes (labeled either 0 or 1) belong to 3 (densely connected) groups with $\sim$ 50 nodes each;
- edges are generated independently with a given probability (high intra-group probability and low inter-group probability)

Compared dissimilarities:

- shortest path length;
- dissimilarity based on the number of common neighbors.

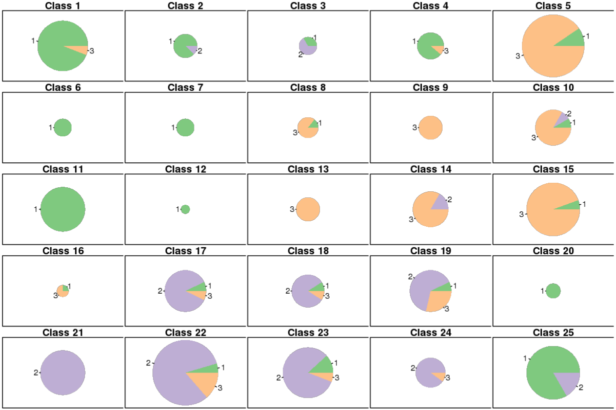
Results



the two distances seem to be approximately equivalent

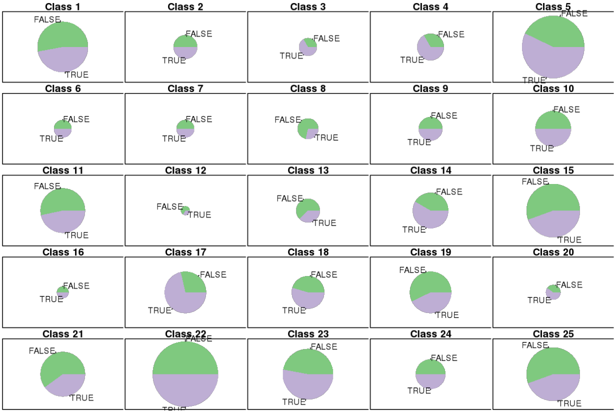
Results

clusters are well organized on the map...



Results

... and clusters contain the two types of nodes



Real-world bipartite graph

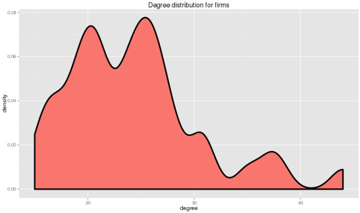
CAC 40

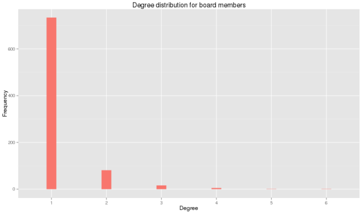
- nodes: CAC 40 firms (40) and board members (838)
- edges: membership (975; *i.e.*, probability that an edge exists between a firm and a board member $\sim$ 2.9%)

Real-world bipartite graph

CAC 40

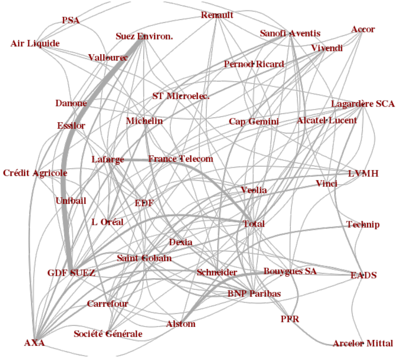
- nodes: CAC 40 firms (40) and board members (838)
- edges: membership (975; *i.e.*, probability that an edge exists between a firm and a board member $\sim$ 2.9%)





- firms have from 15 to 45 board members (most of them have less than 30 board members)
- most board members are involved in only one firm (more than 700), a few board members are involved in up to 6 firms

Firm maps



- interesting facts: central position of Total, position at borders of Arcelor Mittal, EADS, PPR, close positions of Renaud and PSA...
- issues to work on: distant positions of Suez Environnement and of GDF Suez, position at border of AXA...

Conclusion

- methods to cluster and display bipartite graphs;
- work in progress to produce a map of CAC40 firms and board members;
- in development: integration of additional information regarding board members personal information (e.g., studies...) to improve the map

Thank you for your attention...

... questions?

Cottrell, M., Letrémy, P., and Roy, E. (1993).
Analyzing a contingency table with Kohonen maps: a factorial correspondence analysis.
In Cabestany, J., Mary, J., and Prieto, A. E., editors, *Proceedings of International Workshop on Artificial Neural Networks (IWANN 93)*, Lecture Notes in Computer Science, pages 305–311. Springer Verlag.

Goldfarb, L. (1984).
A unified approach to pattern recognition.
Pattern Recognition, 17(5):575–582.

Kohonen, T. (2001).
Self-Organizing Maps, 3rd Edition, volume 30.
Springer, Berlin, Heidelberg, New York.

Olteanu, M. and Villa-Vialaneix, N. (2014).
On-line relational and multiple relational SOM.
Neurocomputing.
Forthcoming.