Hiérarchies, hiérarchies faibles et convexités d'intervalle

Patrice Bertrand[1]

[1]Ceremade, Université Paris-Dauphine, PSL Research, Paris, France

Co-auteur : Jean Diatta
Université de La Réunion, Saint-Denis, France

SFC-2019, 3-5 septembre, Nancy - France

Plan

Plan

Plan

Plan

Plan

Plan

Introduction

Soit S un ensemble fini (ensemble des objets à classer)

- *Convexité*
  Collection de parties de S contenant S et stable par intersections
- *Fonction intervalle*
  $I : S \times S \mapsto 2^S$ tq I est symétrique et $I(x, y)$ contient x, y
- *Convexité induite par une fonction d'intervalle*
  Ensemble des parties C de S qui contiennent $I(x, y)$ si $x, y \in C$
- Hiérarchies, hiérarchies faibles, pyramides ...
  Classifications multi-niveaux vues comme *convexités d'intervalle*
- *Avantages*
  Flexibilité et progressivité de la définition d'une fonction intervalle
- *Inconvénients*
  Peu adapté pour définir une classe à l'aide de son centre

Classifications multi-niveaux

Classifications multi-niveaux

↪ Il existe au moins une classe strictement incluse dans une autre.

Hiérarchies : Correspondance bijective avec les ultramétriques
Johnson (1967), Benzécri (1973)

Extension du modèle hiérarchique

Extensions : recouvrements possibles entre les classes

- **Hiérarchies faibles**
  Bandelt & Dress (1989, 1994), Diatta & Fichet (1994, 1998), ...
- **Classifications pyramidales** (ou pseudo-hierarchies)
  Diday (1984, 1986), Fichet (1984, 1986))
- **Hiérarchies sur paires**
  B. (2002, 2008), B. & Brucker (2007)
- **Hiérarchies *k*-faibles**
  Bandelt & Dress (1994), Diatta (1997), B. & Janowitz (2003)

Définitions

$\mathcal{F} \subseteq 2^S$ est un *système de classification* si $\mathcal{F}$ contient S mais pas $\emptyset$

De plus, $\mathcal{F}$ est dit :

- *hiérarchique*

  si $X \cap Y \in \{\emptyset, X, Y\}$ pour tout $X, Y \in \mathcal{F}$

- *hiérarchique sur paires*

  si $\#\{Y \in \mathcal{F} \setminus X \mid X \cap Y \neq \emptyset\} \leq 1$ pour tout $X \in \mathcal{F}$

- *pyramidal*

  s'il existe un ordre total de S tq chaque membre de $\mathcal{F}$ soit un intervalle

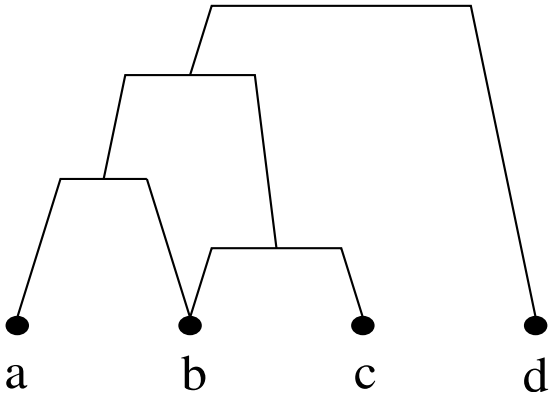
- *faiblement hiérarchique*

  si $X \cap Y \cap Z \in \{X \cap Y, Y \cap Z, X \cap Z\}$ pour tout $X, Y, Z \in \mathcal{F}$

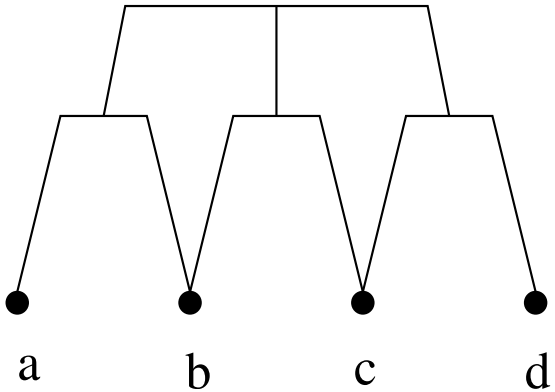
- *k-faiblement hiérarchique*

  si $\bigcap_{i \in [k+1]} X_i \in \left\{ \bigcap_{i \in [k+1] \setminus \{j\}} X_i \mid 1 \leq j \leq k+1 \right\}$ pour tout $X_i, \ldots, X_{k+1} \in \mathcal{F}$

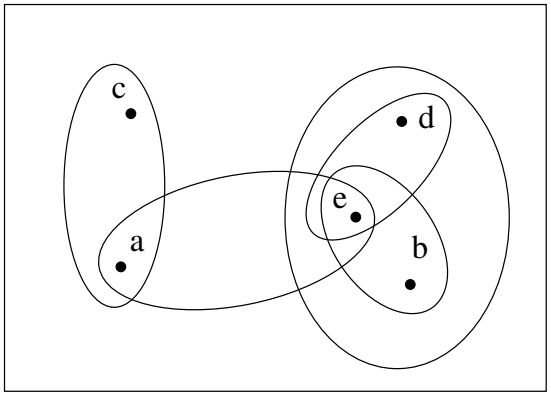
Exemples



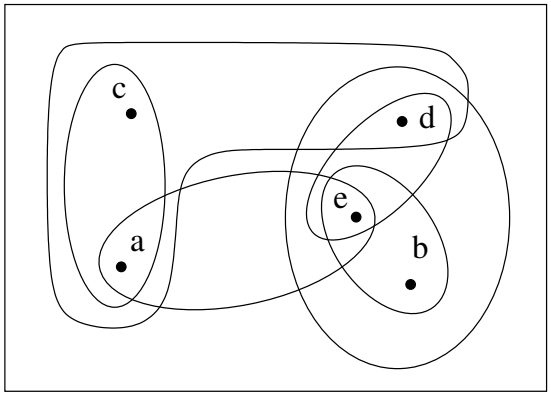
(a) hiérarchie sur paire



(b) classification pyramidale



(c) hiérarchie faible



(d) hiérarchie 3-faible

Convexités abstraites

- Une collection $\mathcal{F} \subseteq 2^S$ est une *convexité* sur S si $\mathcal{F}$ est fermée par intersections, et contient $\emptyset$ et S.
- Les membres de $\mathcal{F}$ sont appelés *ensembles convexes*.
- Si $A \subseteq S$, alors $hull_{\mathscr{C}}(A) := \bigcap\{C \ : \ A \subseteq C \in \mathscr{C}\}$ est l'*enveloppe* de A
- Si $a, b \in S$, alors $\mathrm{seg}_{\mathscr{C}}(a, b) = hull_{\mathscr{C}}(\{a, b\})$ est le *segment entre a et b*
- $I : S^2 \mapsto 2^S$ est une *fonction intervalle sur S* si pour tout $a, b \in S$, on a : $a, b \in I(a, b) = I(b, a)$
- L'ensemble $\mathrm{conv}(I) = \{C \subseteq S \mid I(x, y) \subseteq C$ pour tout $x, y \in C, \}$ est appelé *convexité d'intervalle induite par I*

$\mathcal{C}$aractérisations

Soit $f : S^2 \mapsto 2^S$ une fonction symétrique.

(HI) Pour tout $x, y, z \in S$, $f(x, y) \subseteq f(x, z)$ ou $f(x, z) \subseteq f(x, y)$

(WH) $\nexists\, x_1, x_2, x_3 \in S$ tq $x_i \notin \bigcup f(x_j, x_\ell)$ pour tout $\{i, j, \ell\} = \{1, 2, 3\}$

Soit $\mathscr{C}$ une convexité :

- $\text{seg}_{\mathscr{C}}$ vérifie (HI) $\Leftrightarrow$ $\mathscr{C}$ est hiérarchique
- $\text{seg}_{\mathscr{C}}$ vérifie (WH) $\Leftrightarrow$ $\mathscr{C}$ est faiblement hiérarchique

Soit I une fonction d'intervalle :

- I vérifie (HI) $\Rightarrow$ conv(I) est hiérarchique
- I vérifie (WH) $\Rightarrow$ conv(I) est faiblement hiérarchique

Propriété : $\mathscr{C} = \text{conv}(\text{seg}_{\mathscr{C}})$ si $\mathscr{C}$ est une convexité faiblement hiérarchique

$\mathcal{F}$onctions g, J_g et M_g

Soit $g : S \times S \mapsto 2^S$ qui vérifie :

$$(C_0) \text{ Pour tout } x, y \in S, \{x, y\} \subseteq g(x, y)$$

Afin de symétriser g, posons :

$$\forall x, y \in S, \quad J_g(x, y) = g(x, y) \cup g(y, x) \quad \text{et} \quad M_g(x, y) = g(x, y) \cap g(y, x)$$

Notons :

(H) pour tout $x_1, x_2, x_3 \in S$, on a :
$g(x_1, x_2) \subseteq g(x_1, x_3)$ ou $g(x_1, x_3) \subseteq g(x_1, x_2)$

(W) pour tout $x_1, x_2, x_3 \in S$, il existe i, j, k tels que $\{i, j, k\} = \{1, 2, 3\}$,
$g(x_i, x_j) \subseteq g(x_i, x_k)$ et $g(x_k, x_j) \subseteq g(x_k, x_i)$

(W′) pour tout $x_1, x_2, x_3 \in S$, il existe i, j, k tels que $\{i, j, k\} = \{1, 2, 3\}$,
$g(x_i, x_j) \subseteq g(x_i, x_k)$, $g(x_k, x_j) \subseteq g(x_k, x_i)$ et $g(x_j, x_i) \subseteq g(x_j, x_k)$

Remarque : (W′) $\Rightarrow$ (H) et (W′) $\Rightarrow$ (W), mais (H) $\not\Rightarrow$ (W) et (W) $\not\Rightarrow$ (H)

Propriétés

Proposition 1

- Si g vérifie (H), alors $\mathrm{conv}(J_g)$ est hiérarchique
- Si g vérifie (W), alors $\mathrm{conv}(M_g)$ est faiblement hiérarchique
- Si g vérifie (W$'$), alors $\mathrm{conv}(J_g)$ et $\mathrm{conv}(M_g)$ sont resp. hiérarchique et faiblement hiérarchique

Proposition 2

- g est symétrique $\Leftrightarrow$ $J_g = M_g$ $\Leftrightarrow$ $J_g = M_g = g$
- Si g vérifie (H), et $C \in \mathrm{conv}(J_g)$, alors pour tout $x_0 \in C$, il existe $y_0 \in C$ tel que $C = g(x_0, y_0)$

Hiérarchie d'Asprejan et hiérarchie faible de Bandelt et Dress

Soit δ une dissimilarité sur S, i.e.

$$\delta(x, y) = \delta(y, x) \geq \delta(x, x) = 0 \text{ pour tout } x, y \in S$$

Asprejan (1966), puis Epter et al. (1999) définissent une classe comme toute partie C vérifiant :

$$\text{pour tout } x, y \in C, \quad \delta(x, y) < \min_{z \notin C} \{\delta(x, z), \delta(y, z)\}$$

Bandelt et Dress (1989), et Bandelt (1992) proposent une condition plus faible :

$$\text{pour tout } x, y \in C, \quad \delta(x, y) < \max_{z \notin C} \{\delta(x, z), \delta(y, z)\}$$

Hiérarchie d'Asprejan et hiérarchie faible de Bandelt et Dress (suite)

Pour tout $x, y \in S$ et $\rho \geq 0$, notons :

$$\begin{aligned} B^c(x,\rho) &= \text{boule fermée de centre } x \text{ et de rayon } \rho \text{ au sens de } \delta \\ g_B(x,y) &= B^c(x,\delta(x,y)) = \{s \in S \mid \delta(x,s) \leq \delta(x,y)\} \\ \mathcal{D}(x,y) &= g_B(x,y) \cup g_B(y,x) = \{z \in S \mid \min\{\delta(x,z),\delta(y,z)\} \leq \delta(x,y)\} \\ \mathcal{B}(x,y) &= g_B(x,y) \cap g_B(y,x) = \{z \in S \mid \max\{\delta(x,z),\delta(y,z)\} \leq \delta(x,y)\} \end{aligned}$$

Or $\mathcal{D} = J_{g_B}$, $\mathcal{B} = M_{g_B}$ et g_B vérifie (W′), d'où la propriété (i) suivante.

Proposition 3

(i) $\mathrm{conv}(\mathcal{D})$ est hiérarchique, et $\mathrm{conv}(\mathcal{B})$ est faiblement hiérarchique

(ii) Une partie est une classe d'Asprejan ssi elle est $\mathcal{D}$-convexe.

(iii) Une partie est une classe de Bandelt et Dress ssi elle est $\mathcal{B}$-convexe.

Par la suite, $\mathcal{H}_A(\delta) = \mathrm{conv}(\mathcal{D})$ désigne la hiérarchie d'Asprejan

$\mathcal{F}_{BD}(\delta) = \mathrm{conv}(\mathcal{B})$ désigne la hiérarchie faible de Bandelt et Dress

Hiérarchie d'Asprejan et hiérarchie faible de Bandelt et Dress (suite)

Soient f_1 et f_2 deux fonctions de $S \times S$ dans 2^S. On note :

$f_1 \preceq f_2$ si $f_1(x, y) \subseteq f_2(x, y)$ pour tout $x, y \in S$

$(f_1 \vee f_2)(x, y) = f_1(x, y) \cup f_2(x, y)$, pour tout $x, y \in S$

Proposition 4

(i) $\mathcal{H}_A(\delta) \subseteq \mathcal{H}_{BD}(\delta)$

(ii) Si $I_1 \preceq I_2$, alors $\mathrm{conv}(I_2) \subseteq \mathrm{conv}(I_1)$

Pour tout $\alpha \in [0, 1]$ et $x, y \in S$, notons :

$$\mathcal{D}_\alpha(x, y) = \{z \in S \mid \min\{\delta(x, z), \delta(y, z)\} \leq \alpha\, \delta(x, y)\}$$

Soit $0 < \alpha \leq \beta < 1$ et δ une dissimilarité propre.

$$\mathcal{B} = (\mathcal{B} \vee \mathcal{D}_0) \preceq (\mathcal{B} \vee \mathcal{D}_\alpha) \preceq (\mathcal{B} \vee \mathcal{D}_\beta) \preceq (\mathcal{B} \vee \mathcal{D}_1) = \mathcal{D}$$

D'où :

$$\mathcal{H}_A(\delta) = \mathrm{conv}(\mathcal{D}_1) \subseteq \mathrm{conv}(\mathcal{B} \vee \mathcal{D}_\beta) \subseteq \mathrm{conv}(\mathcal{B} \vee \mathcal{D}_\alpha) \subseteq \mathrm{conv}(\mathcal{B}) = \mathcal{F}_{BD}(\delta)$$

CAH avec le lien simple : notations

Soient :

- $\Gamma[\alpha]$, le graphe seuil (inférieur) de la dissimilarité δ au niveau α
- $K(S)$ le graphe complet défini sur S
- $\ell \in \{1, \ldots, n-1\}$ avec $n = |S|$
- $\mathcal{P}_x^{(\ell)}$ l'ensemble des chemins de K_S de longueur au plus ℓ et dont une extrémité est x. On note $\mathcal{P}_{x-y}^{(\ell)} = \mathcal{P}_x^{(\ell)} \cap \mathcal{P}_y^{(\ell)}$
- Pour tout chemin $P = \{u_0, u_1, \ldots, u_m\}$ de K_S et $x, y \in S$, on note :

$$\pi^{\ell}(x,y) = \min_{P \in \mathcal{P}_{x-y}^{(\ell)}} \Big[\max_{1 \le i \le m} \delta(u_{i-1}, u_i) \Big]$$

- Soit l'application g^{ℓ} de $S \times S$ dans 2^S, définie par :

$$g^{\ell}(x,y) = \bigcup \{ P \in \mathcal{P}_x^{(\ell)} \mid \max_{1 \le i \le m} \delta(u_{i-1}, u_i) \le \pi^{\ell}(x,y) \}$$

Il en résulte que $x, y \in g^{\ell}(x,y)$. On pose :

$$\begin{aligned} J^{\ell}(x,y) &= J_{g^{\ell}}(x,y) = g^{\ell}(x,y) \cup g^{\ell}(y,x), \\ M^{\ell}(x,y) &= M_{g^{\ell}}(x,y) = g^{\ell}(x,y) \cap g^{\ell}(y,x) \end{aligned}$$

$\mathcal{F}$onctions g^ℓ, J^ℓ et M^ℓ

Par la suite, $\mathcal{H}_S(\delta)$ désigne la hiérarchie du lien simple

Proposition 5

Soient $A \subseteq S$ et $\ell \in \{1, \ldots, n-1\}$.

(i) g^ℓ vérifie la condition (W′)

(ii) g^{n-1} est symétrique

(iii) A est J^ℓ-convexe ssi il existe $\alpha \geq 0$ telle que A est une composante connexe de $\Gamma[\alpha]$ avec $\operatorname{diam}_{\Gamma[\alpha]}(A) \leq \ell$

Corollaire

(i) $\operatorname{conv}(M^1)$ coincide avec $\mathcal{F}_{BD}(\delta)$

(ii) Pour tout $1 \leq \ell_1 \leq \ell_2 \leq n-1$, on a :

$$\mathcal{H}_A(\delta) = \operatorname{conv}(J^1) \subseteq \operatorname{conv} J^{\ell_1} \subseteq \operatorname{conv} J^{\ell_2} \subseteq \operatorname{conv}(J^{n-1}) = \mathcal{H}_S(\delta)$$

Propriétés

Proposition 6

- $\mathcal{H}_A(\delta) \subseteq \mathcal{H}_{SL}(\delta) \cap \mathcal{F}_{BD}(\delta)$
- Les conditions suivantes sont équivalentes :
  (a) δ est une ultramétrique
  (b) $\mathcal{H}_A(\delta) = \mathcal{H}_{SL}(\delta) = \mathcal{F}_{BD}(\delta)$
  (c) $\mathcal{H}_A(\delta) = \mathcal{H}_{SL}(\delta)$

Remarque. Comme le montre le contre-exemple suivant, $\mathcal{H}_{SL}(\delta)$ n'est en général pas incluse dans $\mathcal{F}_{BD}(\delta)$. Soient $S = abcd$ et δ définie par :

δ	b	c	d
a	5	4	5
b		3	6
c			3

Pour ces données,$\mathcal{H}_S(\delta)$ ne contient qu'une seule classe non triviale qui est $C = bcd$. Par ailleurs :

$$\delta(b,d) = 6 \geq \max_{z \notin C}\{\delta(b,z), \delta(d,z)\} = \max\{\delta(b,a), \delta(d,a)\} = 5,$$

Lien complet

On rappelle que la méthode de la C.A.H. utilisant le lien complet, peut générer plusieurs hiérarchies distinctes

Par la suite, on note $\mathcal{H}_{CL}(\delta)$ l'une quelconque des hiérarchies construites par la CAH avec le lien complet

Théorème

La hiérarchie d'Asprejan est incluse dans toute hiérarchie construite, selon la méthode du lien complet, par l'algorithme de la C.A.H. appliqué à la dissimilarité δ

Remarque

- $\mathcal{H}_{CL}(\delta)$ n'est en général pas incluse dans $\mathcal{F}_{BD}(\delta)$
- $\mathcal{H}_{A}(\delta)$ peut être strictement incluse dans l'intersection de toutes les hiérarchies construites par la CAH avec le lien complet

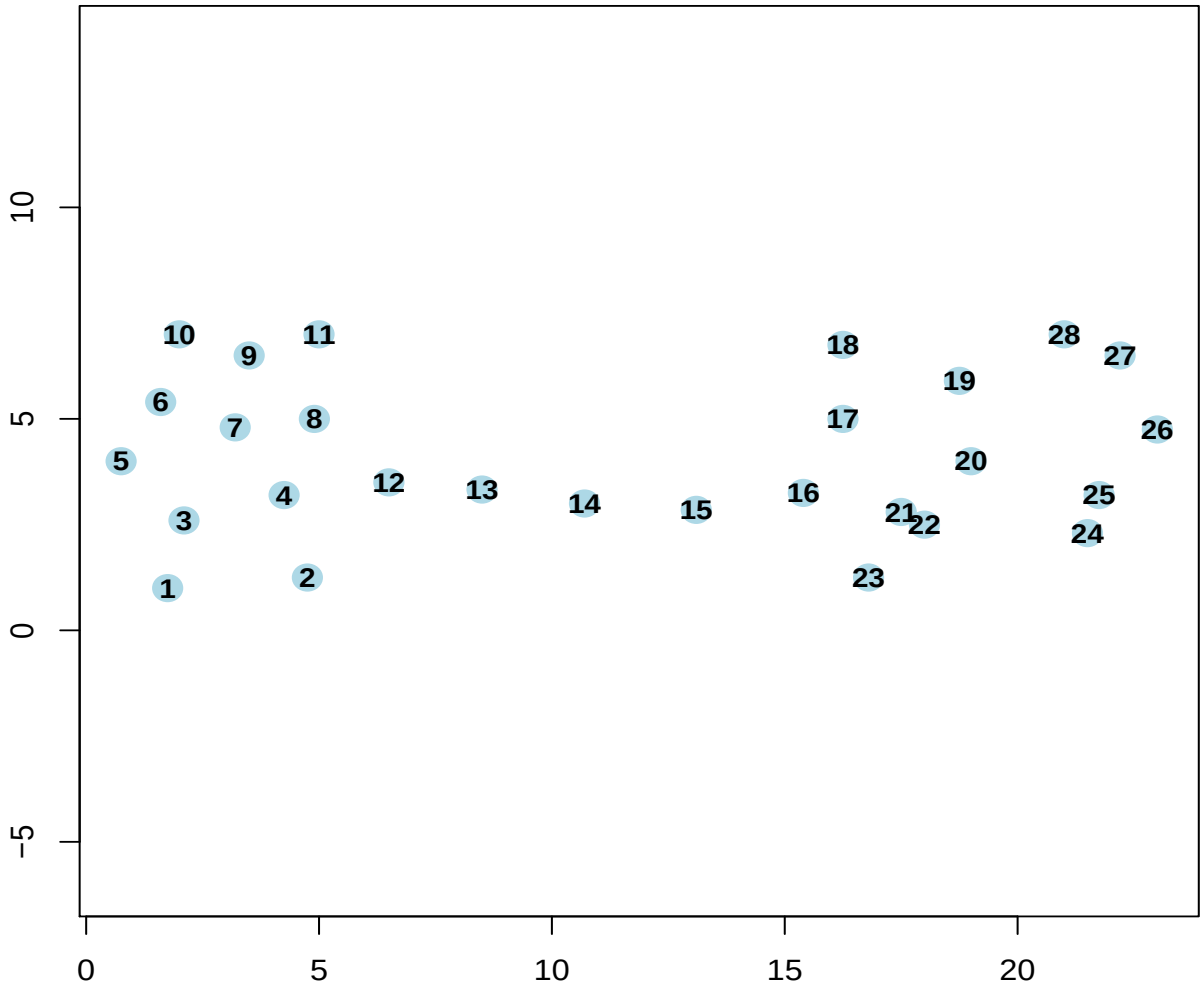
Propriété

Proposition 7

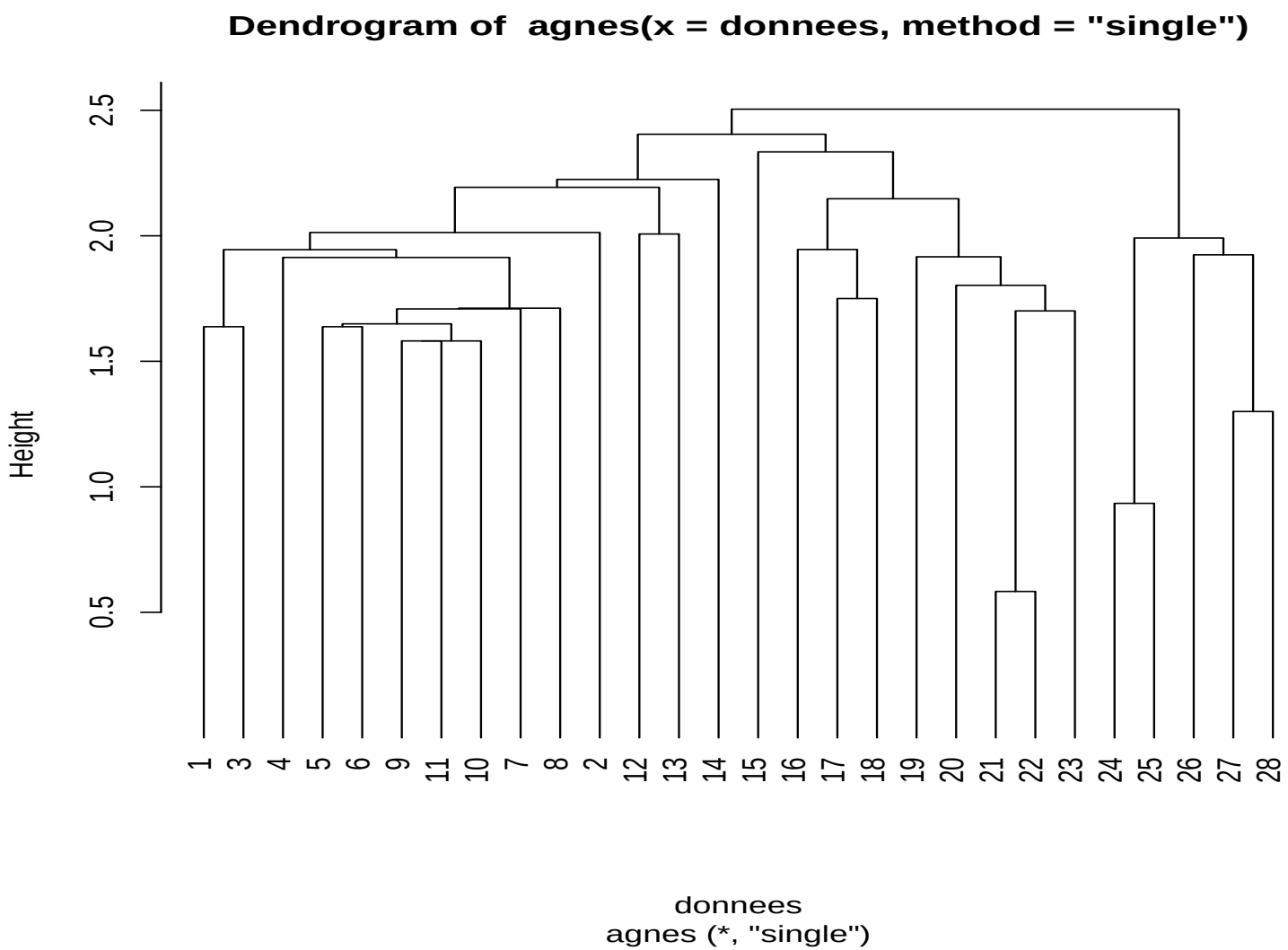
Soient $\mathcal{H}$ une hiérarchie définie sur S et σ la fonction segment associée.
Si $\mathcal{H}_A(\delta) \subseteq \mathcal{H}$, alors :

- La convexité induite par la fonction $\sigma \vee \mathcal{D}$ est égale à $\mathcal{H}_A(\delta)$
- Soit $(\alpha_n)_{0 \leq n \leq N}$ une suite décroissante de réels telle que $\alpha_0 = 1$ et $\alpha_N = 0$, et soit δ une dissimilarité propre.
  La suite $\{\text{conv}(\sigma \vee \mathcal{D}_{\alpha_n})\}_{0 \leq n \leq N}$ est une suite croissante (au sens de l'inclusion) de hiérarchies emboîtées, qui commence avec la hiérarchie d'Asprejan (pour $n = 0$), et se termine avec la hiérarchie $\mathcal{H}$ (pour $n = N$)

Exemple


10
5
0
−5
0
5
10
15
20
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28

CAH avec le lien simple



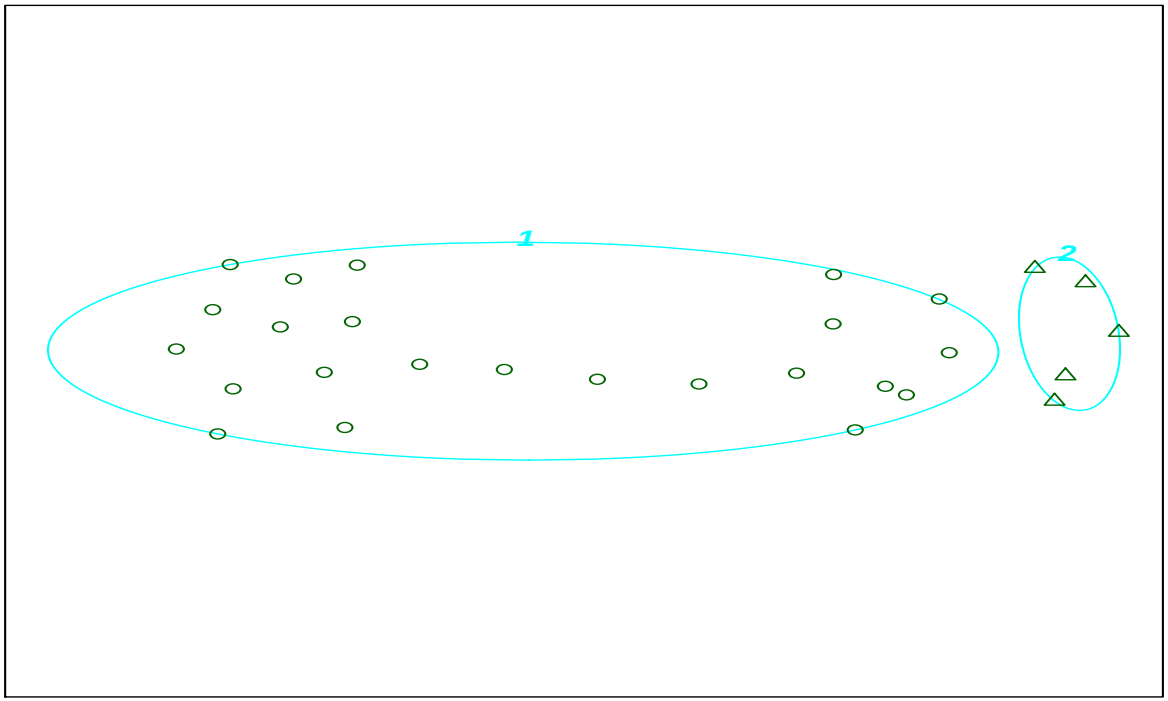
Hiérarchie d'Asprejan

Classes au sens d'Asprejan :

$$\mathcal{H}_A(\delta) = \big\{\{1,3\},\{17,18\},\{24,25\},\{21,22,23\},\{27,28\}\big\} \cup \mathcal{T},$$

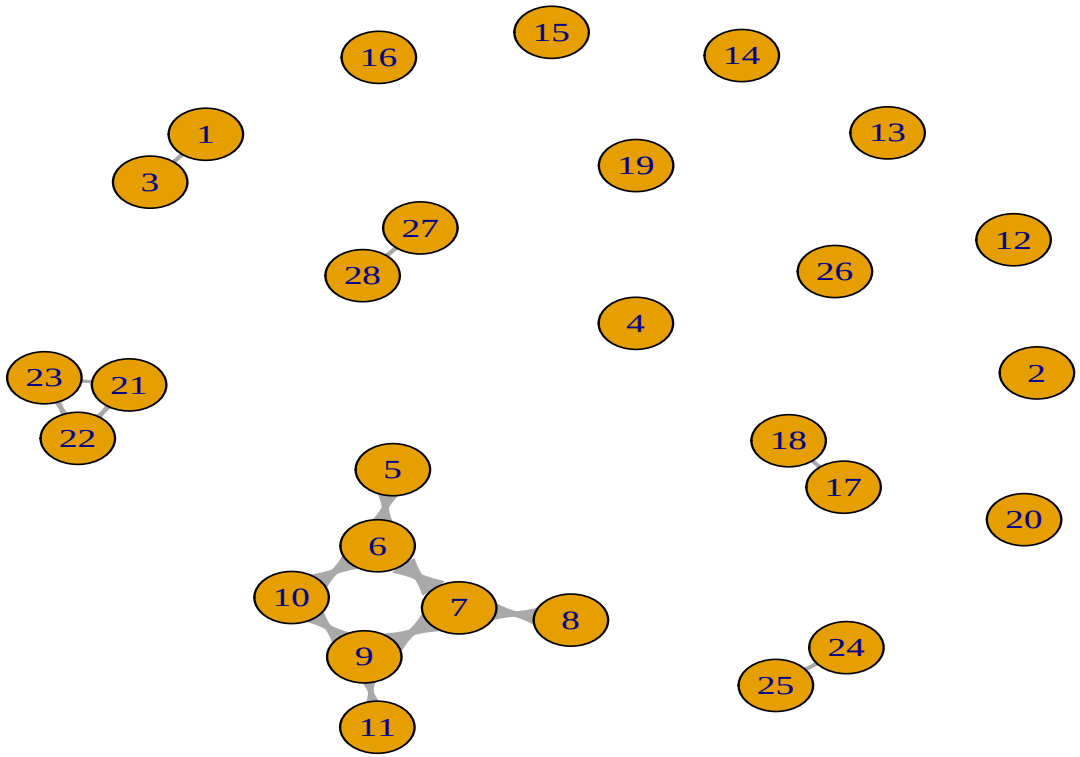
où $\mathcal{T}$ désigne l'ensemble des classes triviales, i.e. $S = \{1,\ldots,28\}$ et ses singletons.

Coupure du dendrogramme en 2 classes



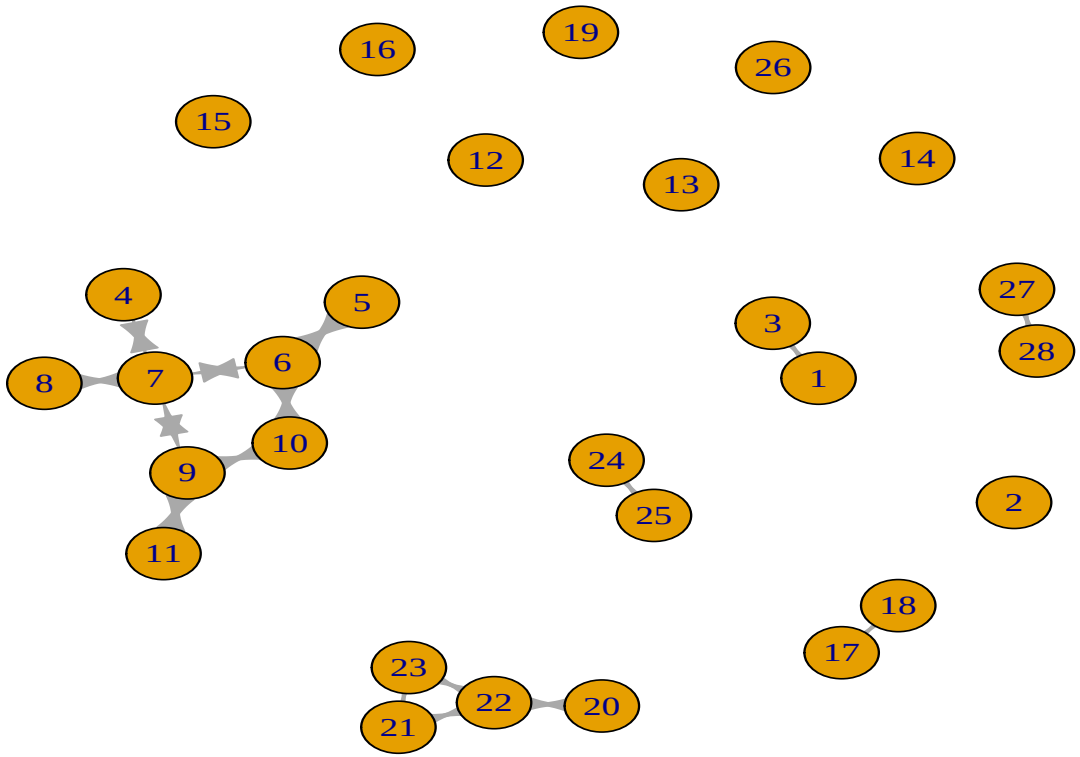
Graphe seuil au 11 ème niveau

Classe de diamètre 4 : $\{5, 6, \ldots, 11\}$



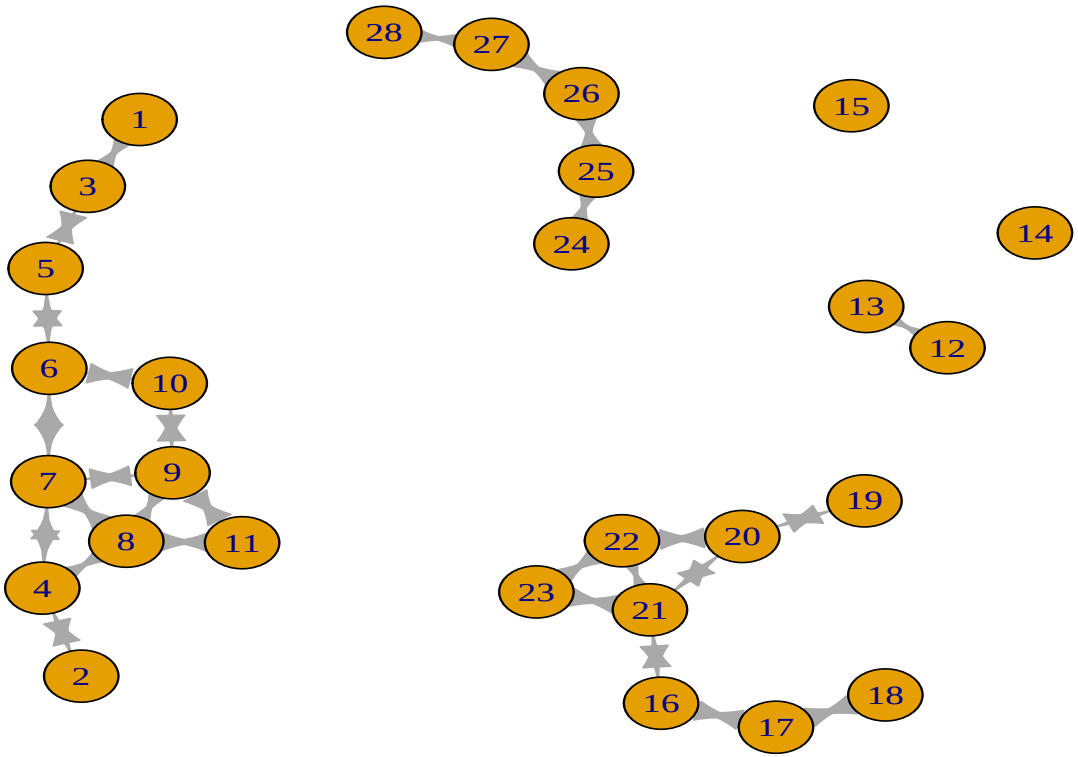
Graphe seuil au 13 ème niveau

Classe de diamètre 4 : $\{4, 5, 6, \ldots, 11\}$



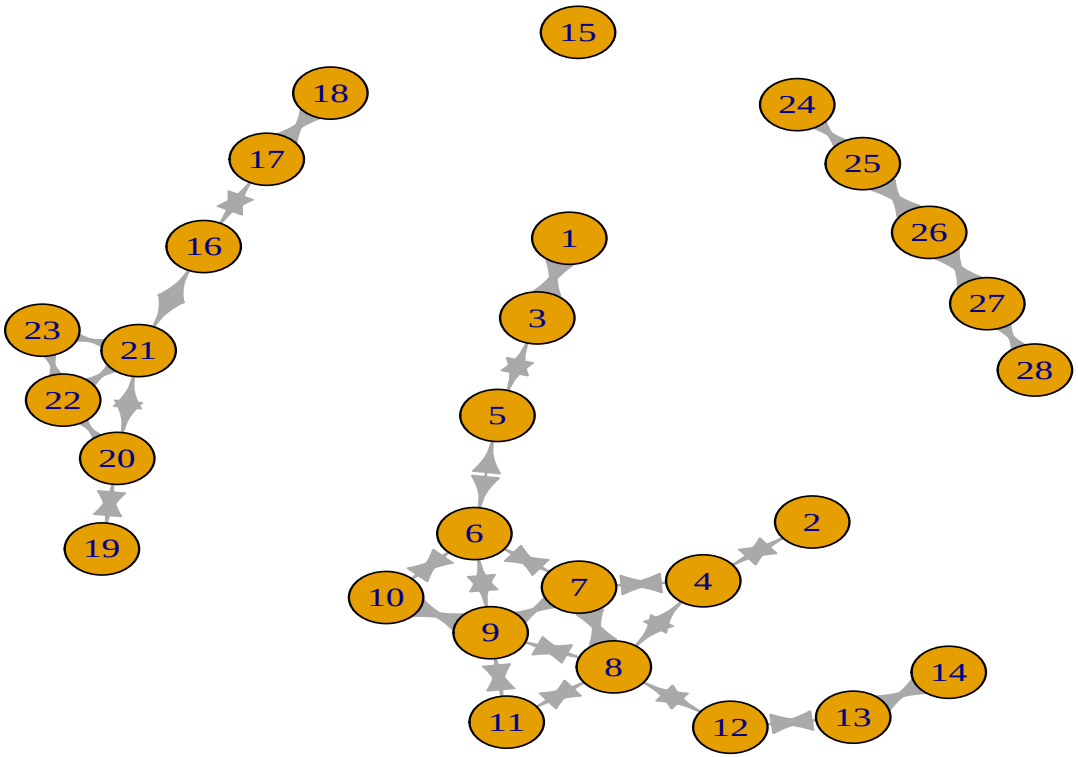
Graphe seuil au 21 ème niveau

Classe de diamètre 6 : $\{1, 2, \ldots, 11\}$



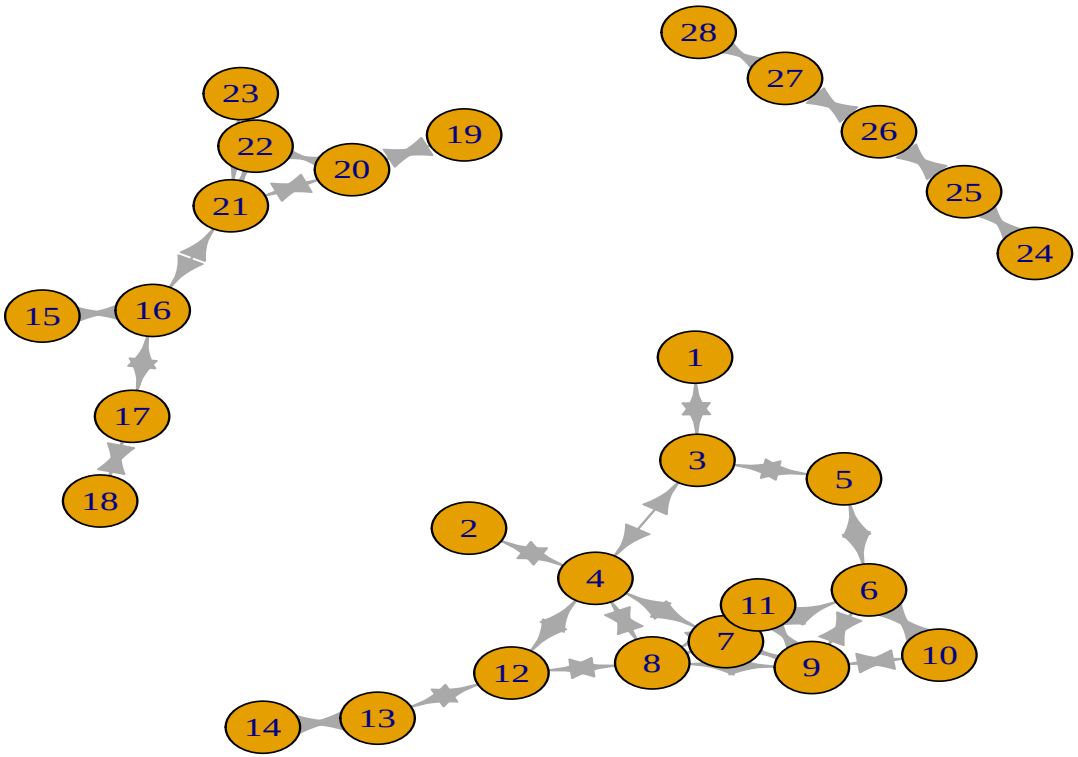
Graphe seuil au 23 ème niveau

Classe de diamètre 8 : $\{1, 2 \ldots, 14\}$



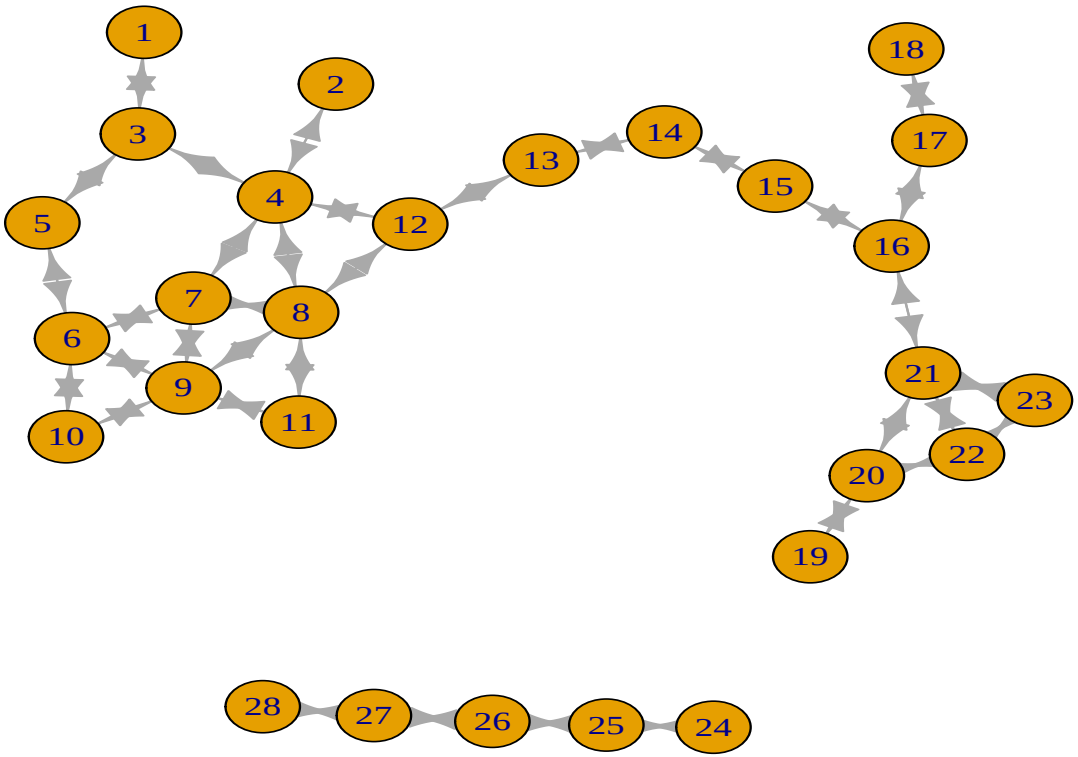
Graphe seuil au 24 ème niveau

Classe de diamètre 5 : $\{1, 2 \ldots, 14\}$



Graphe seuil au 25 ème niveau

Classe de diamètre 10 : $\{1, 2 \ldots, 23\}$



Conclusion

- Procédé de construction de fonctions d'intervalle générant soit une hiérarchie soit une hiérarchie faible
- Suites de hiérarchies emboîtées entre la hiérarchie d'Asprejan et des classifications bien connues (liens simple et complet)
- Suite de hiérarchies faibles emboîtées entre la hiérarchie d'Asprejan et la hiérarchie faible de Bandelt et Dress
- Aides à l'interprétation des structures de classifications contenant la hiérarchie d'Asprejan