Détection de communautés dans les réseaux sociaux

Ricco Rakotomalala

Un réseau social désigne un ensemble d'individus (au sens large : famille, collègues, organisations, etc.) qui sont liés d'une certaine manière, qui interagissent en s'échangeant du contenu. Le domaine a connu une très forte accélération ces dernières années avec le développement des médias sociaux. On parle de réseautage social.

« The top 25 Social Networking Sites People Are Using », E. Moreau, Lifewire, October 2016.

Une communauté est formée par un ensemble d'individus qui interagissent (interaction sociale) plus souvent entre eux qu'avec les autres. Ils s'agit donc de groupes d'individus qui ont tissés des liens plus forts ou qui ont des affinités communes.

C'est une forme de « clustering »

La détection de communautés a pour rôle de mettre en évidence ces groupes qui se sont formés implicitement c.-à-d. qui ne résultent pas d'un choix explicite. Ex. Dans une entreprise, les employés forment des communautés explicites (les services), mais certains vont collaborer plus souvent avec d'autres, dans le service, ou à cheval sur différents services, et former des communautés implicites.

L'intérêt de la détection de communautés est multiple : identifier des profils types, effectuer des actions ciblées, mieux ajuster les recommandations, réorganisation, identifier les acteurs centraux / influents, etc.

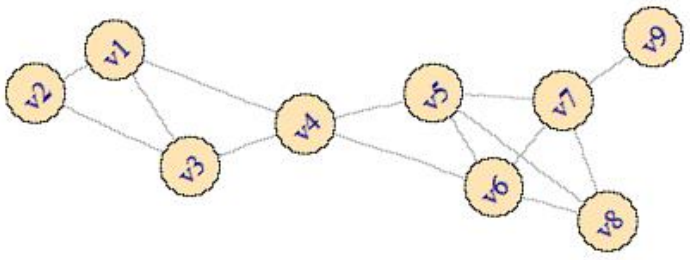
1. Représentation des réseaux sociaux
2. Détection de communautés

Approches hiérarchiques

3. Approche agglomérative (CAH)
4. Approche divisive
5. Approche par maximisation de la modularité
6. Changement de représentation (MDS)
7. Plus loin avec la découverte de communautés
8. Bibliographie

Utilisation des graphes – Matrice d'adjacence

REPRÉSENTATION DES RÉSEAUX SOCIAUX



(Tang & Liu, 2010)

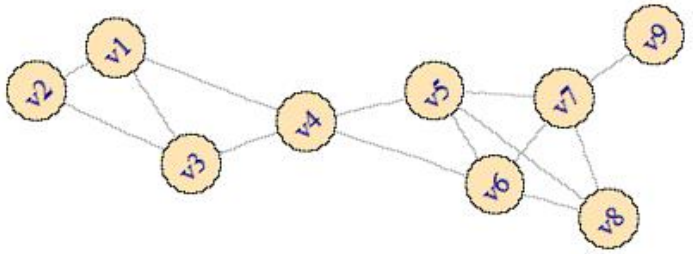
Un réseau social peut être représenté par un graphe G(V,E) où V (*vertex*) représente l'ensemble des sommets (nœuds), E l'ensemble des arêtes (*edge*).

S	v1	v2	v3	v4	v5	v6	v7	v8	v9
v1	0	1	1	1	0	0	0	0	0
v2	1	0	1	0	0	0	0	0	0
v3	1	1	0	1	0	0	0	0	0
v4	1	0	1	0	1	1	0	0	0
v5	0	0	0	1	0	1	1	1	0
v6	0	0	0	1	1	0	1	1	0
v7	0	0	0	0	1	1	0	1	1
v8	0	0	0	0	1	1	1	0	0
v9	0	0	0	0	0	0	1	0	0

Un graphe peut être représenté à l'aide matrice dite d'adjacence (A_{ij}) qui indique les connexions entre les nœuds.

Nous nous positionnons dans un cadre très simple. Le graphe peut être plus complexe, il peut être orienté (arcs entre les sommets), les liens peuvent être pondérés, les sommets peuvent l'être également. Ex. Les citations bibliographiques.

G(V,E) : V l'ensemble des sommets [v_i, i = 1, …, n], E l'ensemble des arêtes [$e_{ij} = e(v_i,v_j)$, i,j = 1,…, n]



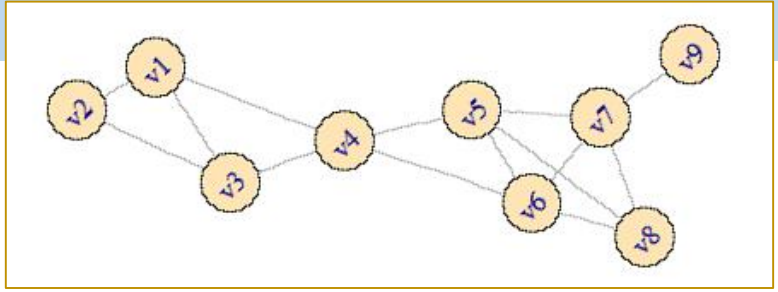
Chemin : Un chemin permettant d'aller de v_i à v_j est la succession de sommets et arêtes qui les deux sommets. De v_i à v_j, plusieurs chemins sont possibles, il existe un plus court chemin minimisant le nombre d'arêtes parcourus.

Boucle : Arête reliant un sommet à lui-même. Nous considérons que nous n'en avons pas.

Distance géodésique δ_{ij} : Nombre d'arêtes du plus court chemin reliant deux sommets. Ex. $\delta_{3,6} = 2$

Voisinage N_i d'un sommet = l'ensemble des nœuds qui lui sont directement connectés. Ex. $N_3 = \{2, 1, 4\}$

Graphe connexe : Il existe au moins un chemin entre chaque paire de nœuds. Il n'y a pas de sommets isolés.



Diamètre d'un graphe : La plus grande distance géodésique possible entre 2 sommets dans le graphe. Diamètre = 5 avec $\delta_{2,9}$

Densité d'un graphe : Rapport entre le nombre d'arêtes observées et le nombre maximal d'arêtes possibles. Densité = 0, tous les sommets sont isolés ; Densité = 1, graphe complet c.-à-d. il y a un lien entre chaque paire de sommets.

$$\text{Densité} = \frac{|E|}{\frac{|V|.(|V|-1)}{2}}$$

Force d'un lien entre deux sommets peut être évalué à partir de l'importance du recouvrement (overlap) entre leurs voisinages. Plus il (le recouvrement) est élevé, plus le lien est fort.

$$\text{overlap}(v_i, vj) = \frac{|N_i \cap Nj|}{|N_i \cup N_j| - 2}$$

Voisinage commun

Voisinage « total »

Ne pas prendre en compte les 2 sommets

$$\text{overlap}(v_4, v_1) = \frac{|\{1,3,5,6\} \cap \{2,3,4\}|}{|\{1,3,5,6\} \cup \{2,3,4\}| - 2} = \frac{1}{6-2} = 0.25$$

$$\text{overlap}(v_4, v_5) = \frac{|\{1,3,5,6\} \cap \{4,6,7,8\}|}{|\{1,3,5,6\} \cup \{4,6,7,8\}| - 2} = \frac{1}{7-2} = 0.2$$

v_1 et v_5 sont tous deux connectés à v_4, mais le lien entre v_4 et v_1 est plus fort (une connexion dans un espace « sparse » est plus importante que dans une zone dense).

Degré de centralité d'un sommet, en terme de voisinage = nombre d'arêtes qui lui est associé (nombre de voisins adjacents), absolu $C_D(v_i)$ ou relatif $C'_D(v_i)$. Ex. $C_D(v_3)= 3$, $C'_D(v_3)= 3/8$

$$C_D(v_i)= d_i = \sum_j A_{ij}$$

$$C'_D(v_i) = \frac{C_D(v_i)}{n-1}$$

S	v1	v2	v3	v4	v5	v6	v7	v8	v9
v1	0	1	1	1	0	0	0	0	0
v2	1	0	1	0	0	0	0	0	0
v3	1	1	0	1	0	0	0	0	0
v4	1	0	1	0	1	1	0	0	0
v5	0	0	0	1	0	1	1	1	0
v6	0	0	0	1	1	0	1	1	0
v7	0	0	0	0	1	1	0	1	1
v8	0	0	0	0	1	1	1	0	0
v9	0	0	0	0	0	0	1	0	0

Matrice d'adjacence

Autre vision de la centralité d'un sommet, en terme de distance : dans quelle mesure le sommet permet d'accéder rapidement aux autres nœuds

$$C_{avg}(4) = 1.63$$
$$C_{avg}(3) = 2.13$$

	v1	v2	v3	v4	v5	v6	v7	v8	v9
v1	0	1	1	1	2	2	3	3	4
v2	1	0	1	2	3	3	4	4	5
v3	1	1	0	1	2	2	3	3	4
v4	1	2	1	0	1	1	2	2	3
v5	2	3	2	1	0	1	1	1	2
v6	2	3	2	1	1	0	1	1	2
v7	3	4	3	2	1	1	0	1	1
v8	3	4	3	2	1	1	1	0	2
v9	4	5	4	3	2	2	1	2	0

Matrice des distances géodésiques

Centralité moyenne : moyenne des distances du nœud à tous les autres (valeur faible = nœud central)

$$C_{avg}(v_i) = \frac{1}{n-1}\sum_{j\neq i} \delta_{ij}$$

Closeness centrality : inverse du degré moyen (valeur forte = nœud central)

$$C_C(v_i) = \frac{n-1}{\sum_{j\neq i} \delta_{ij}}$$

Betweenness centrality : recenser le plus court chemin entre chaque paire de nœuds, comptabiliser le nombre de fois où il passe par le sommet étudié (rapporté sur le nombre de plus courts chemins possibles entre ces nœuds – voir illustration edge betweenness). Plus le *betweenness centrality* est grand, plus le sommet est central puisqu'il est situé à la « croisée » des chemins.

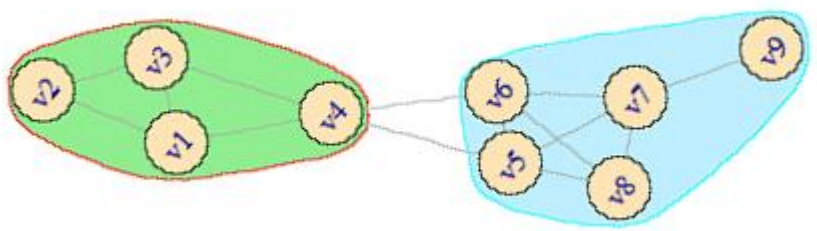
Définition et principe

DÉTECTION DE COMMUNAUTÉS

Une communauté est un ensemble de personnes qui entretiennent des liens privilégiés parce ce qu'ils ont des affinités particulières, ou présentent des caractéristiques similaires, ou encore partagent des centres d'intérêts, etc.

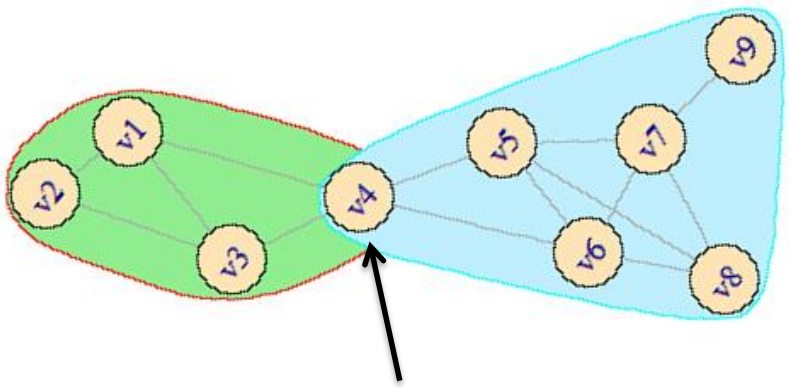
Au sens du graphe, une communauté est constitué par un ensemble de nœuds qui sont fortement liés entre eux, et faiblement liés avec les nœuds situés en dehors de la communauté.

Une partition possible de notre réseau exemple.



C1 C2

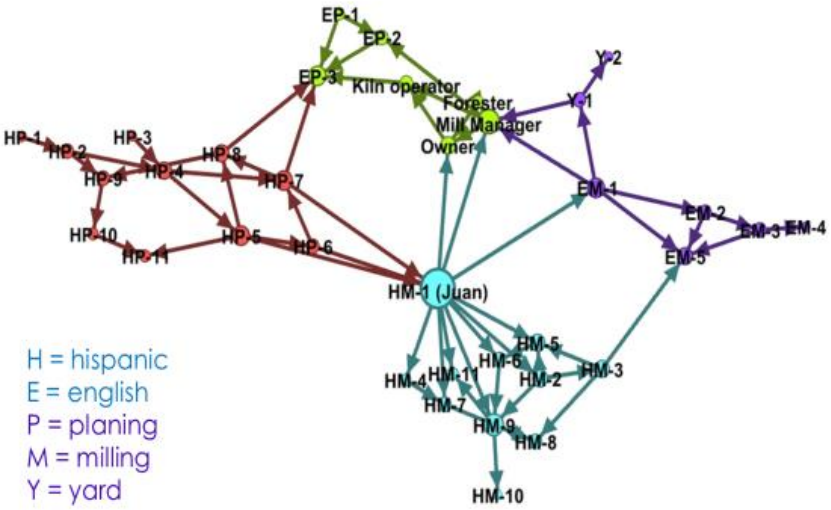
Exactement comme en classification automatique (clustering), la partition peut être « crisp » (groupes disjoints) ou « floue » (groupes avec chevauchement). L'appartenance à une communauté n'est pas forcément univoque.



L'individu n°4 fait partie des deux communautés, il joue le rôle d'interface entre les deux. On perçoit la notion de centralité et d'influence.

Nous nous intéressons aux communautés disjointes dans ce support.

Exemple de la « scierie de bois » (Michael & Massey, 1997).



(Selmane, 2015)

Les employés étaient de groupes ethniques différents (E, H), et affectés à différentes tâches (P, M, Y). Lorsqu'un changement d'organisation a été proposé, une opposition forte est apparue.

Une cartographique des communications (fréquence des discussions) a été réalisée. Il s'est avéré que JUAN tenait un rôle particulier dans la structure. Il fallait le persuader lui d'abord, l'acceptation se propagerait aux autres par la suite. C'est ce qui s'est passé.

Force d'une communauté : écart entre le nombre de connexions observées et le nombre théorique (espéré) obtenu si les connexions avaient été distribuées aléatoirement entre les nœuds (Q élevé ➔ communauté compacte)

Espérance de la connexion sous hypothèse de distribution aléatoire.

$$Q(C) = \sum_{i,j \epsilon C} A_{ij} - \frac{d_i d_j}{2m}$$

m est le nombre de connexions observées

d_i,d_j sont les degrés de centralité des nœuds v_i et v_j

C est la communauté étudiée.

Modularité d'un réseau après partitionnement en K communautés.

Voir https://en.wikipedia.org/wiki/Modularity_(networks)

$$Q = \frac{1}{2m} \sum_{k=1}^{K} \sum_{i,j \epsilon Ck} \left(A_{ij} - \frac{d_i dj}{2m} \right)$$

Plus la modularité est grande, meilleure est la partition.

On peut écrire :

$$Q = \frac{1}{2m} Tr(S^T B S)$$

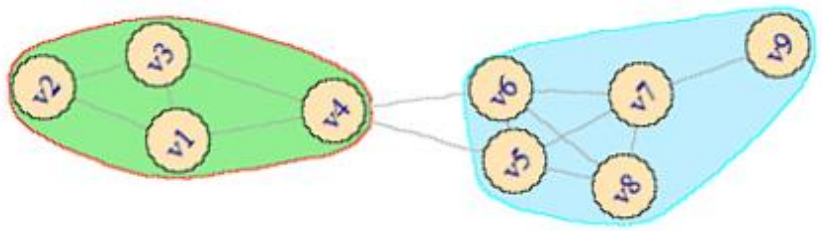
Où m est le nombre de connexions dans le réseau

S (n x K) est une matrice d'indicatrices associant chaque individu à une des communautés.

B (n x n) est la matrice de modularité

$$B_{ij} = A_{ij} - \frac{d_i d_j}{2m}$$

m=14



S

	C1	C2
V1	1	0
V2	1	0
V3	1	0
V4	1	0
V5	0	1
V6	0	1
V7	0	1
V8	0	1
V9	0	1

B

	V1	V2	V3	V4	V5	V6	V7	V8	V9
V1	-0.321	0.786	0.679	0.571	-0.429	-0.429	-0.429	-0.321	-0.107
V2	0.786	-0.143	0.786	-0.286	-0.286	-0.286	-0.286	-0.214	-0.071
V3	0.679	0.786	-0.321	0.571	-0.429	-0.429	-0.429	-0.321	-0.107
V4	0.571	-0.286	0.571	-0.571	0.429	0.429	-0.571	-0.429	-0.143
V5	-0.429	-0.286	-0.429	0.429	-0.571	0.429	0.429	0.571	-0.143
V6	-0.429	-0.286	-0.429	0.429	0.429	-0.571	0.429	0.571	-0.143
V7	-0.429	-0.286	-0.429	-0.571	0.429	0.429	-0.571	0.571	0.857
V8	-0.321	-0.214	-0.321	-0.429	0.571	0.571	0.571	-0.321	-0.107
V9	-0.107	-0.071	-0.107	-0.143	-0.143	-0.143	0.857	-0.107	-0.036

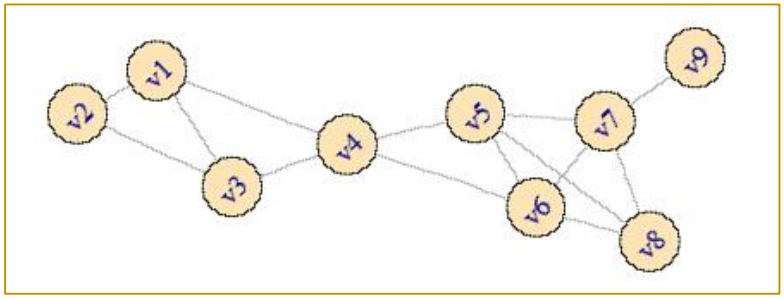
$$Q = 0.347$$

Peut servir à comparer deux partitions concurrentes, ou à détecter le bon nombre de partitions (faire varier K et surveiller l'évolution de Q, comme en clustering).

Algorithme pour la détection des communautés

APPROCHE AGGLOMÉRATIVE

Principe : On définit une mesure de similarité entre les sommets à partir de la matrice d'adjacence, et nous pouvons enchaîner avec une CAH (Ex. saut moyen, etc.)



Indice de Jaccard $Jaccard(v_i, v_j) = \frac{|N_i \cap N_j|}{|N_i \cup N_j|}$

$$\text{Jaccard}(v_4, v_5) = \frac{|\{1,3,5,6\} \cap \{4,6,7,8\}|}{|\{1,3,5,6\} \cup \{4,6,7,8\}|} = \frac{1}{7} = 0.143$$

Similarité Cosinus $Cosine(v_i, v_j) = \frac{|N_i \cap N_j|}{\sqrt{|N_i| . |N_j|}}$

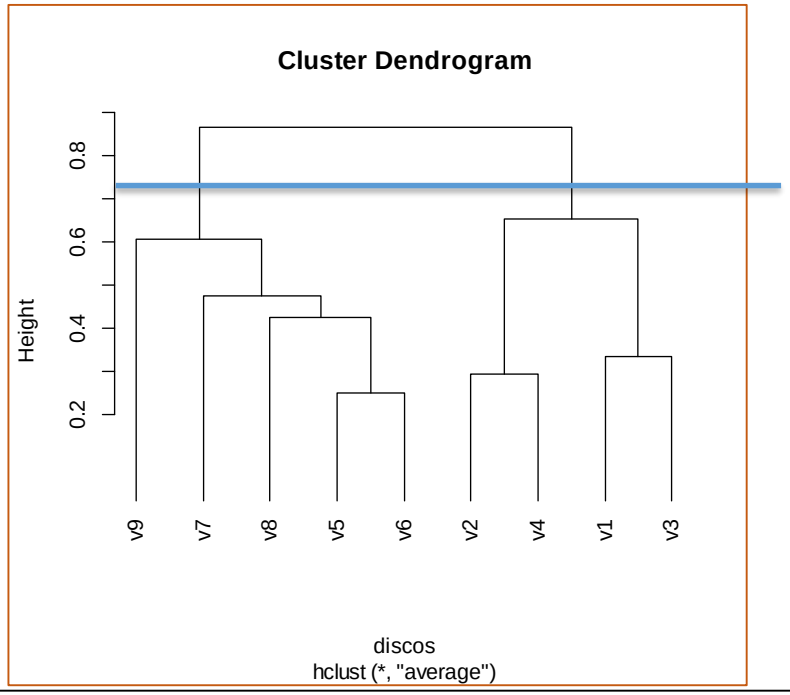
$$\text{Cosine}(v_4, v_5) = \frac{|\{1,3,5,6\} \cap \{4,6,7,8\}|}{\sqrt{|\{1,3,5,6\}| . |\{4,6,7,8\}|}} = \frac{1}{4} = 0.25$$

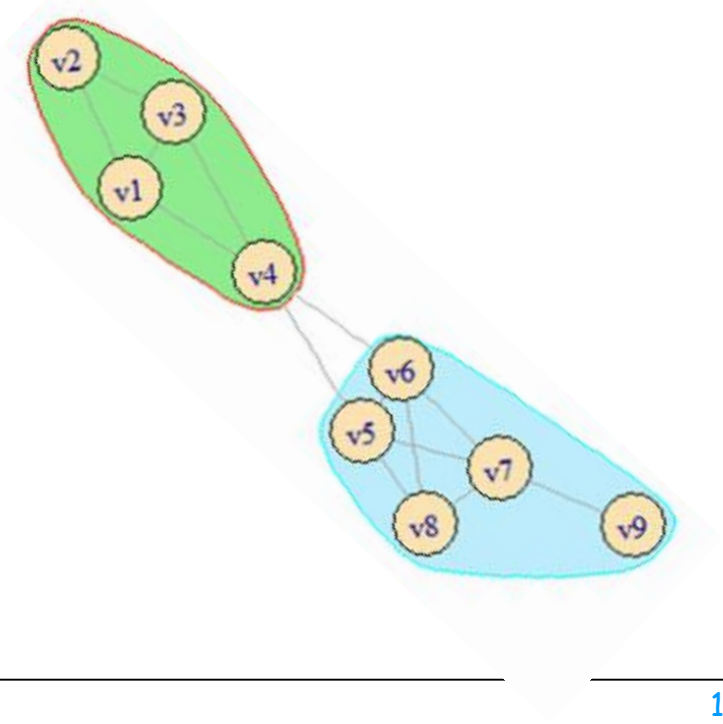
La similarité est définie en terme de voisinage : deux sommets sont « proches » s'il y a un fort recouvrement entre leurs voisinages.

CAH avec dissimilarité = (1 – cos) + Méthode du saut moyen.

Matrice de similarité Cosinus

	V1	V2	V3	V4	V5	V6	V7	V8	V9
v1	1	0.408	0.667	0.289	0.289	0.289	0.000	0.000	0.000
v2	0.408	1	0.408	0.707	0.000	0.000	0.000	0.000	0.000
v3	0.667	0.408	1	0.289	0.289	0.289	0.000	0.000	0.000
v4	0.289	0.707	0.289	1	0.250	0.250	0.500	0.577	0.000
v5	0.289	0.000	0.289	0.250	1	0.750	0.500	0.577	0.500
v6	0.289	0.000	0.289	0.250	0.750	1	0.500	0.577	0.500
v7	0.000	0.000	0.000	0.500	0.500	0.500	1	0.577	0.000
v8	0.000	0.000	0.000	0.577	0.577	0.577	0.577	1	0.577
v9	0.000	0.000	0.000	0.000	0.500	0.500	0.000	0.577	1





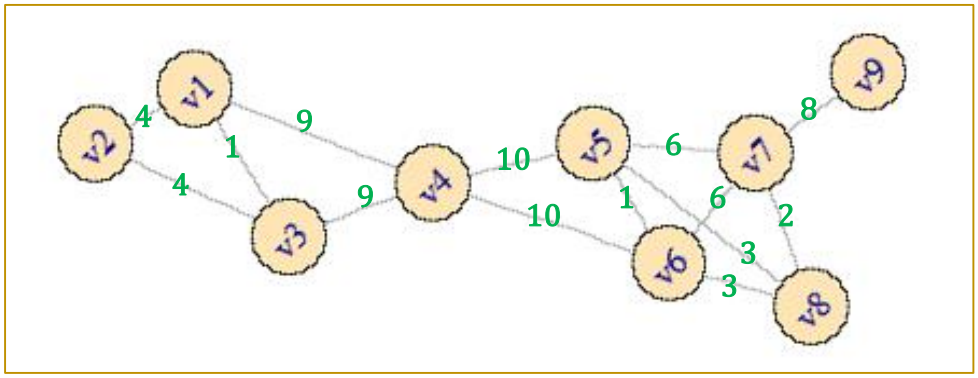
APPROCHE DIVISIVE

L'importance d'une connexion entre deux sommets peut être matérialisée par le « edge betweenness ». Elle indique la fréquence avec laquelle elle est empruntée lorsque l'on considère le plus court chemin entre chaque paire de nœuds. Plus la valeur est élevée, plus la connexion est importante parce qu'elle établit un « pont » entre des groupes de sommets.

$$eb(e) = \sum_{i} \sum_{j>i} \frac{\sigma_{ij}(e)}{\sigma_{ij}}$$

Nombre de plus courts chemins entre les sommets v_i et v_j passant par la connexion e

Nombre de plus courts chemins entre les sommets v_i et v_j



eb(e[1,2]) = 4, parce que…

1-> (2,3,…) $\frac{1}{1} + \frac{0}{1} + \frac{0}{1} + \cdots$

2-> (3,4,…) $\frac{0}{1} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}$

3-> (4,5,…) $\frac{0}{1} + \frac{0}{1} + \frac{0}{1} + \frac{0}{2} + \cdots$

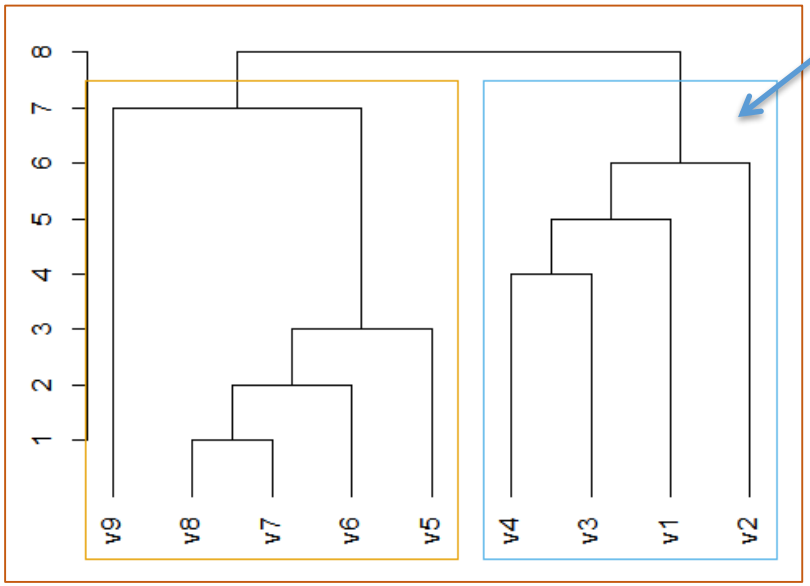
• • •

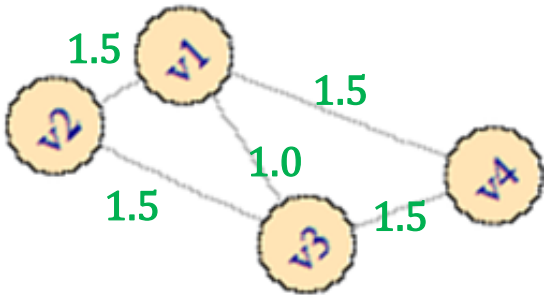
$\Sigma = 4$

La procédure repose sur un partitionnement récursif

```
Fonction Subdiviser(graphe)
Calculer les eb()
TANT QUE graphe ≠ singleton
  Retirer le lien avec l'eb() le plus élevé
  Si partition en g1 et g2 Alors
     Subdiviser(g1), Subdiviser(g2)
  Sinon
     Recalculer les eb()
FIN TANT QUE
```

Dans les sous-graphes, les « edge betweenness » doivent être recalculés.





On obtient un « dendrogramme » du graphe

APPROCHE PAR MAXIMISATION DE LA MODULARITÉ

La modularité Q permet d'évaluer la qualité d'une partition. Pourquoi ne pas la maximiser explicitement en cherchant la matrice d'appartenance S qui permet de l'optimiser ?

$$\max_S Q = \frac{1}{2m} Tr(S^T B S)$$

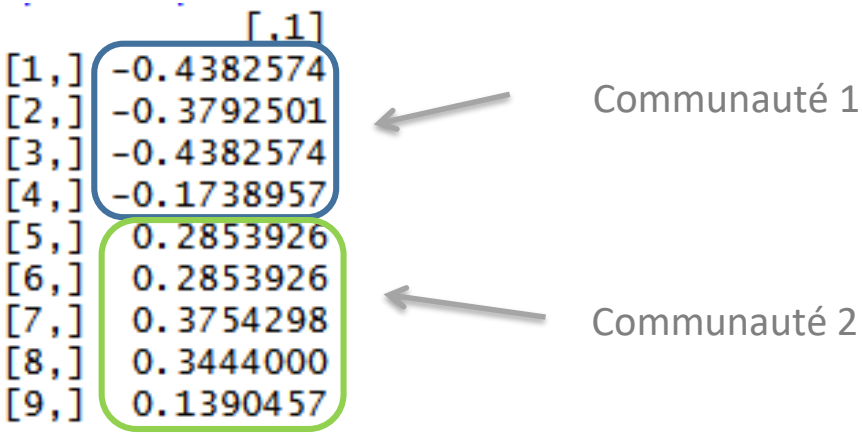
w.r.t $S^T S = I$

S n'est plus une matrice d'indicatrices fournie, mais devient le fruit du calcul. Nous avons une sorte de repère factoriel où les axes sont deux à deux orthogonaux

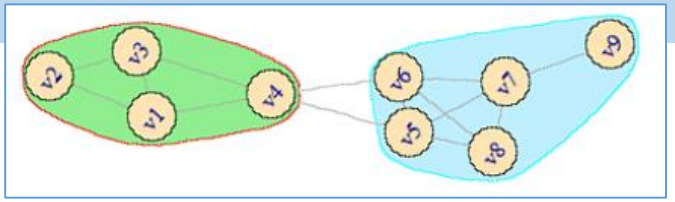
- Les solutions sont obtenues grâce à la diagonalisation de la matrice de modularité B (écart entre les matrices d'adjacence observées et théoriques)
- Seuls les facteurs correspondants aux valeurs propres positives contribuent à la modularité.
- Les vecteurs propres associés permettent de positionner les sommets dans un repère factoriel.

On s'en tient à la première valeur propre.

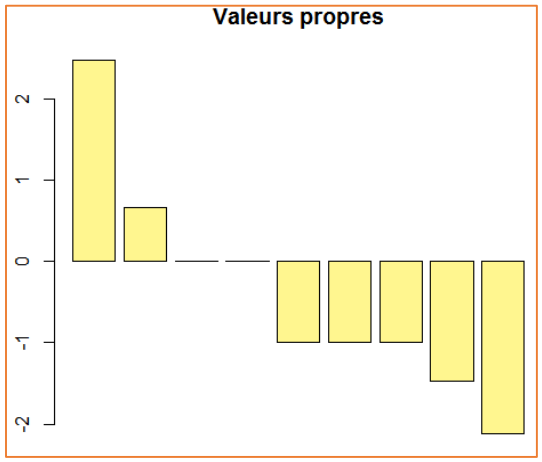
$diag(B)$



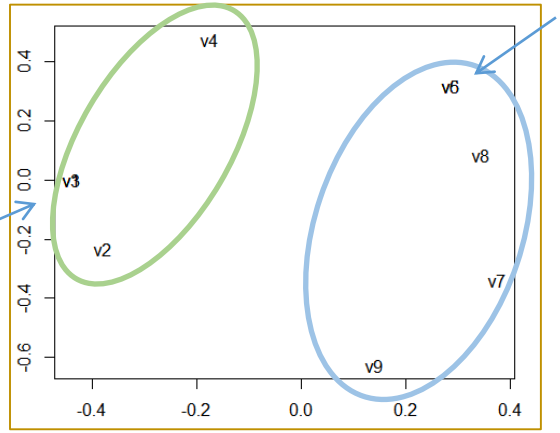
- On subdivise en deux groupes en fonction du signe des coefficients du vecteur propre.
- S'ils ont tous le même signe, cela veut dire qu'il n'y a pas de partition possible.
- Pour une subdivision en K (K > 2) communautés, il faut soit adopter une approche divisive, soit voir la page suivante.



$diag(B)$



2 valeurs propres positives



V5 et V6 sont confondus

V1 et V3 sont confondus

On positionne les sommets dans ce nouvel espace

Et on peut alors utiliser un algorithme classique de clustering (ex. K-Means)

Placer les sommets dans une espace factoriel et réaliser un clustering

POSITIONNEMENT MULTIDIMENSIONNEL

(MDS EN ANGLAIS)

Le MDS est une technique d'analyse factorielle sur matrice de distances. On souhaite que la distance entre chaque paire d'individu soit préservée au mieux dans le repère factoriel.

On peut utiliser la matrice des distances géodésiques

$$\Delta = (\delta_{ij})\ ;\ i,j = 1,\dots,n$$

On cherche à minimiser le critère

$$\min \sum_{i,j}^{n} (\theta_{ij} - \delta_{ij})^2$$

θ_{ij} est la distance entre les sommets v_i et v_j dans le nouvel espace de représentation défini par les coordonnées factorielles.

Pour cela on doit diagonaliser la matrice Δ' avec

$$\delta'_{ij} = -\frac{1}{2}\left(\delta_{ij}^2 - \delta_{i.}^2 - \delta_{.j}^2 + \delta_{..}^2\right) \quad \text{où} \quad \begin{cases} \delta_{i.}^2 = \frac{1}{n}\sum_j \delta_{ij}^2 \\ \delta_{.j}^2 = \frac{1}{n}\sum_i \delta_{ij}^2 \\ \delta_{..}^2 = \frac{1}{n^2}\sum_{i,j} \delta_{ij}^2 \end{cases}$$

Les coordonnées factorielles s'écrivent alors

$$F = V\Lambda^{\frac{1}{2}}$$

Λ est la matrice diagonale des plus grandes valeurs propres (positives)
V sont les vecteurs propres corresp.

2 premières v.p.

$\Delta' =$

	v1	v2	v3	v4	v5	v6	v7	v8	v9
v1	2.457	3.957	1.957	0.846	-0.654	-0.654	-2.210	-2.043	-3.654
v2	3.957	6.457	3.957	1.346	-1.154	-1.154	-3.710	-3.543	-6.154
v3	1.957	3.957	2.457	0.846	-0.654	-0.654	-2.210	-2.043	-3.654
v4	0.846	1.346	0.846	0.235	-0.265	-0.265	-0.821	-0.654	-1.265
v5	-0.654	-1.154	-0.654	-0.265	0.235	-0.265	0.679	0.846	1.235
v6	-0.654	-1.154	-0.654	-0.265	-0.265	0.235	0.679	0.846	1.235
v7	-2.210	-3.710	-2.210	-0.821	0.679	0.679	2.123	1.790	3.679
v8	-2.043	-3.543	-2.043	-0.654	0.846	0.846	1.790	2.457	2.346
v9	-3.654	-6.154	-3.654	-1.265	1.235	1.235	3.679	2.346	6.235

$\Lambda =$

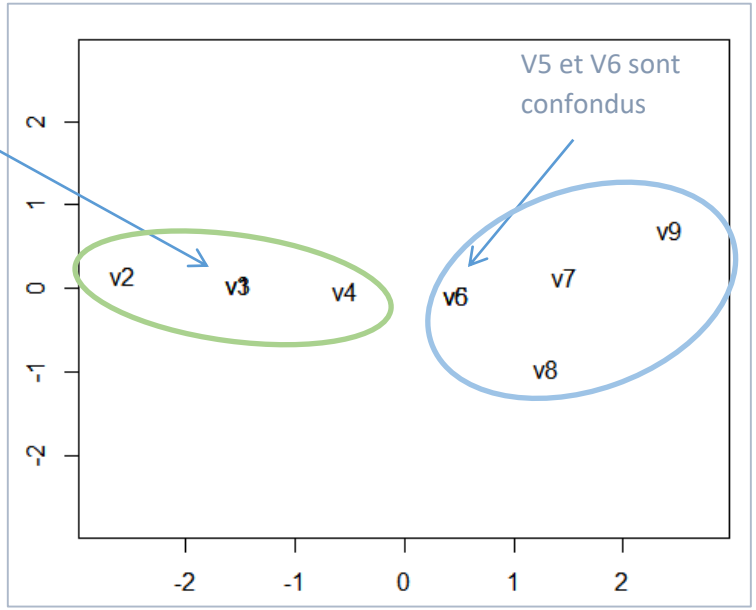
```
          [,1]     [,2]
[1,] 21.55732 0.000000
[2,]  0.00000 1.462843
```

$V =$

```
            [,1]         [,2]
[1,] -0.3258546  0.047207013
[2,] -0.5519664  0.136651635
[3,] -0.3258546  0.047207013
[4,] -0.1148586 -0.009966884
[5,]  0.1013447 -0.062063001
[6,]  0.1013447 -0.062063001
[7,]  0.3157864  0.112895783
[8,]  0.2786439 -0.788417450
[9,]  0.5214146  0.578548893
```

V1 et V3 sont confondus

V5 et V6 sont confondus



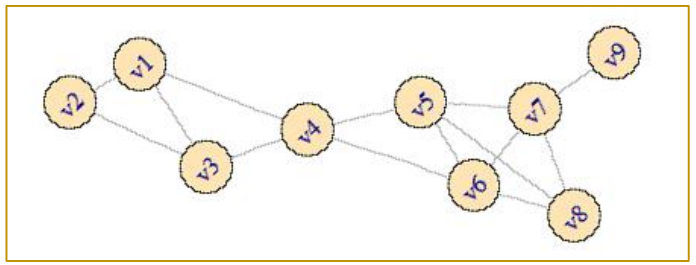

```
            [,1]       [,2]
[1,] -1.5129384  0.05709596
[2,] -2.5627721  0.16527750
[3,] -1.5129384  0.05709596
[4,] -0.5332868 -0.01205475
[5,]  0.4705422 -0.07506399
[6,]  0.4705422 -0.07506399
[7,]  1.4661916  0.13654525
[8,]  1.2937394 -0.95357558
[9,]  2.4209204  0.69974364
```

$F = V\Lambda^{\frac{1}{2}}$

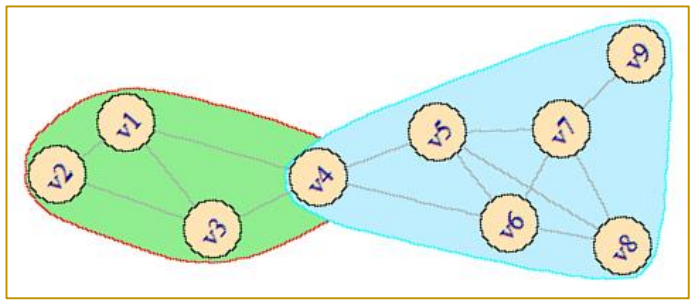
On peut appliquer un algorithme de clustering dans le nouvel espace de représentation.

PLUS LOIN AVEC LA DÉCOUVERTE DE COMMUNAUTÉS

Une autre approche possible de la découverte de communautés est l'étude des cliques. Une clique est un sous-ensemble de sommets qui sont en relation deux à deux. On parle de clique maximale lorsqu'elle n'est pas contenue dans une clique plus grande. Ex. {5, 6, 7} est une clique ; {5, 6, 7, 8} est une clique maximale.



Déjà évoquée plus haut, la recherche des communautés chevauchantes permet d'enrichir l'analyse car plus réaliste. On tient compte de la multi-appartenance, un facteur d'appartenance est calculé. Ex. un personne peut aimer le tennis ET le foot.



La recherche des influences et l'étude de la propagation des informations sont des domaines particulièrement intéressants. On travaille à partir d'un graphe orienté, on étudie l'activation des sommets et sa diffusion dans le graphe. Ex. étude de la propagation des fake news.

Les réseaux sociaux sont souvent hétérogènes. Ex. Twitter. Les entités (sommets) peuvent être de type différents (les utilisateurs, les hashtags, les URL), et les relations (liens entre les entités) peuvent s'appuyer sur des supports de différents types (échanges en ligne, mail personnel, followers, etc.). L'intégration de ces différents modes et dimensions enrichit l'analyse mais la complexifie également.

L'étude des communautés dans les réseaux sociaux dynamiques est un champ d'application très intéressant. Les liens ne sont pas statiques, les acteurs évoluent, les structures des communautés également : elles peuvent disparaître, fusionner, certaines peuvent émerger, etc. On ajoute une dimension temporelle dans l'analyse.

Voir notre site de référence : http://dmml.asu.edu/cdm/

Ouvrages et articles

- Tang L., Liu H., « Community detection and mining in social media », Morgan and Claypool Publishers, 2010. **A servi de référence**, une bonne partie du matériel pédagogique est accessible en ligne : http://dmml.asu.edu/cdm/
- Selmane S.A., « Détection et analyse de communautés dans les réseaux sociaux : approche basée sur l'analyse formelle de concepts », Thèse de Doctorat Lyon 2, 2015.
- Newman M.E.J., « Finding community structure in networks using eigenvectors of matrices », Physical Review, E 74 036104, 2006.
- Package igraph, « The network analysis ».
- Liu B., « Web Data Mining - Exploring Hyperlinks, Contents, and Usage Data », Springer, 2008 ; Chapitre 7.
- Leskovec J., Rajaraman A., Ullman J., « Mining of Massive Datasets », 2nd Edition, 2014. Chapitre 10.
- Wikipédia, « Community structure ».