Algorithmes de graphes pour l'étude de réseaux sociaux

Mikaela Keller

June 28, 2019

Des réseaux un peu partout

On entend parler de réseaux dans de nombreux contextes:

- réseau ferroviaire
- réseau social
- réseau internet
- réseau de neurones
- ...

Des réseaux un peu partout (2)

Tous ont en commun d'avoir:

des entités:

- des gares
- des gens
- des ordinateurs
- les neurones

des liens entre les entités

- la voie de chemin de fer
- les intéractions
- les cables
- les connections synaptiques

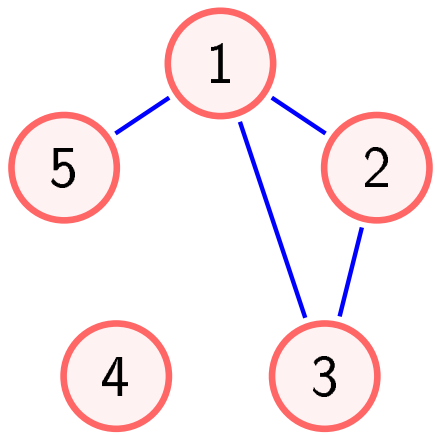
Plan

- Définitions formelles et typologie
- Caractérisation quantitatives
- Visualisation
- Partitionnement

Définition: graphe

Un graphe $G = (V, E)$ est définit par:

- l'ensemble V de ses sommets (*or vertices in English*): càd les entités, les noeuds
- l'ensemble E des arêtes (*or edges in English*): càd les liens entre les entités
- Sur l'exemple:



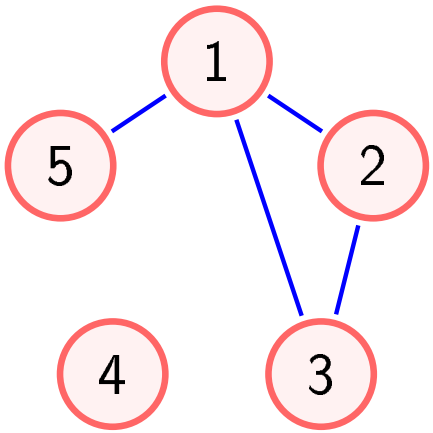
$$V = \{1, 2, 3, 4, 5\}$$

$$E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 5\}\}$$

Définition: matrice d'adjacence

La matrice d'adjacence d'un graphe $G = (V, E)$ est une représentation de G par un tableau rempli de 0 et de 1.

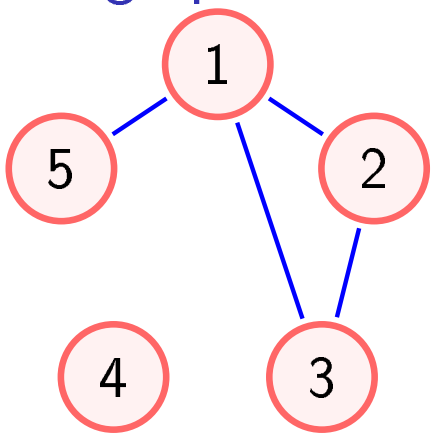
- autant de lignes et de colonnes que de sommets dans V
- un 1 signifie qu'il y a une arête entre les sommets correspondants
- exemple de graphe avec sa matrice d'adjacence:



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Définition: graphe non-dirigé

Un graphe non-dirigé



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

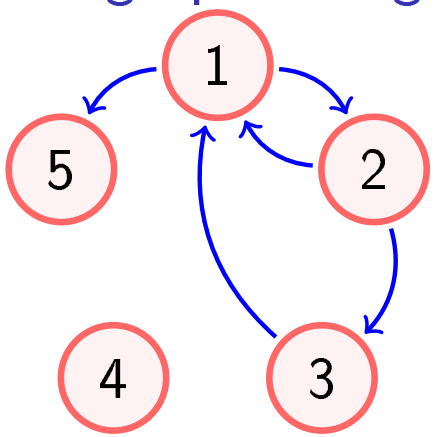
Un graphe $G = (V, E)$ est non-dirigé:

- si la relation entre les entités (*ie* les arêtes) est non directionnelle ou symmétrique
- notation ensembliste: Une arête est un ensemble de 2 éléments dont l'ordre est arbitraire

$$\{1, 2\} = \{2, 1\}$$

Définition: graphe dirigé

Un graphe dirigé



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Un graphe $G = (V, E)$ est dirigé:

- si la relation entre les entités (*ie* les arêtes) est directionnelle
- notation séquentielle: Une arête est un couple de 2 éléments ordonnés: la source, la cible

$$(1, 2) \neq (2, 1)$$

Sur l'exemple:

$$V = \{1, 2, 3, 4, 5\}$$

$$E = \{(1, 2), (2, 1), (1, 3), (2, 3), (1, 5)\}$$

Exemples

Graphes non-dirigés

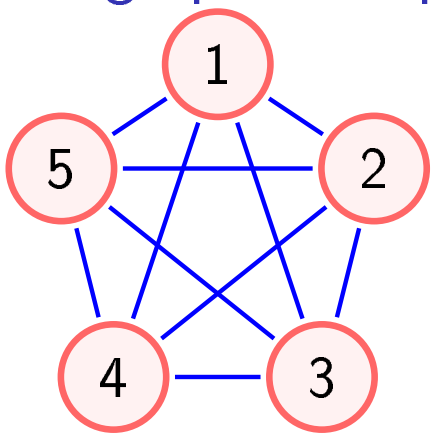
- réseau ferroviaire
- réseau internet
- réseau de collaboration scientifique
- réseaux sociaux virtualisés Facebook, linkedIn

Graphes dirigés

- réseau de neurones
- graphe de citations
- arbre généalogique
- réseaux sociaux virtualisés Twitter, Instagram

Définition: graphe complet

Un graphe complet



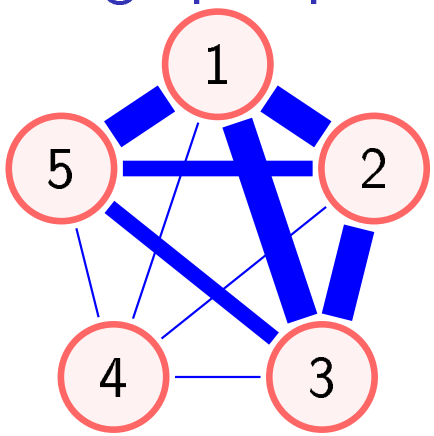
$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Un graphe $G = (V, E)$ est complet: si chaque sommet est relié par une arête aux autres sommets

Un sous-graphe complet s'appelle aussi une clique.

Définition: graphe pondéré

Un graphe pondéré



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & .6 & .6 & .1 & .6 \\ .6 & 0 & .6 & .1 & .3 \\ .6 & .6 & 0 & .1 & .3 \\ .1 & .1 & .1 & 0 & .1 \\ .6 & .3 & .3 & .1 & 0 \end{pmatrix} \end{array}$$

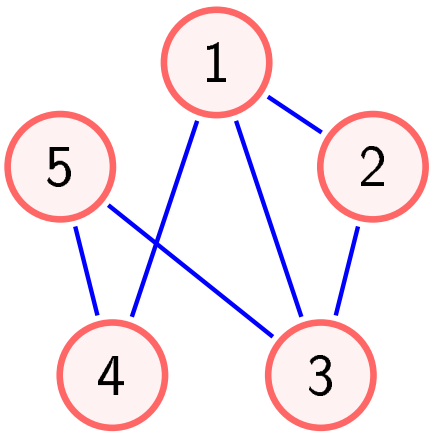
Un graphe $G = (V, E)$ est pondéré s'il a des poids associés à ses arêtes.
On s'intéresse alors à sa matrice de pondération.

Exemples: graphe de similarité

- A des entités $\{x_1, \ldots, x_n\}$ sont associés k caractéristiques
- Ces caractéristiques permettent de calculer une similarité entre entités
- A partir de cette relation de similarité on peut construire un graphe de similarité
- Ce graphe est pondéré et complet par défaut

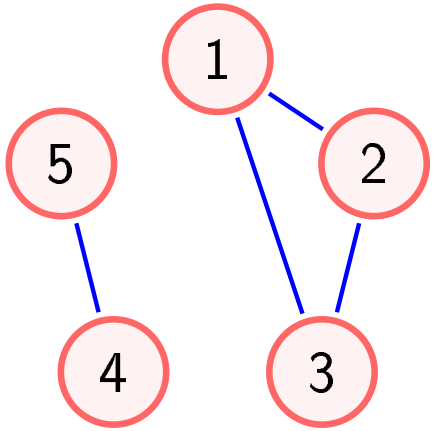
Définition: chemin

- Un chemin est une séquence d'arêtes reliant 2 sommets
- La longueur d'un chemin est définie par le nombre d'arêtes de la séquence
- Chemins entre les sommets 1 et 3
  - 1 - 3 : longueur = 1
  - 1 - 2 - 3 : longueur = 2
  - 1 - 4 - 5 - 3 : longueur = 3



Définition: graphe connexe

Un graphe connexe ?



$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{array}$$

Un graphe $G = (V, E)$ est connexe:

- si pour chaque couple de sommets, il existe un chemin qui les relie
- s'il est non connexe: il est composé de plusieurs composantes connexes
- chaque composante connexe est un sous-graphe connexe

Mesures: Degré - Densité

- le degré d'un noeud est le nombre d'arêtes qui s'attachent à ce noeud
- Soit $G = (V, E)$ un graphe avec n sommets $V = \{x_1, \ldots, x_n\}$ et m arêtes, alors:

$$d(x_1) + \ldots + d(x_n) = 2 \times m$$

- Si G est un graphe complet tous les sommets ont un degré de $n - 1$ alors

$$d(x_1) + \ldots + d(x_n) = n \times (n - 1)$$

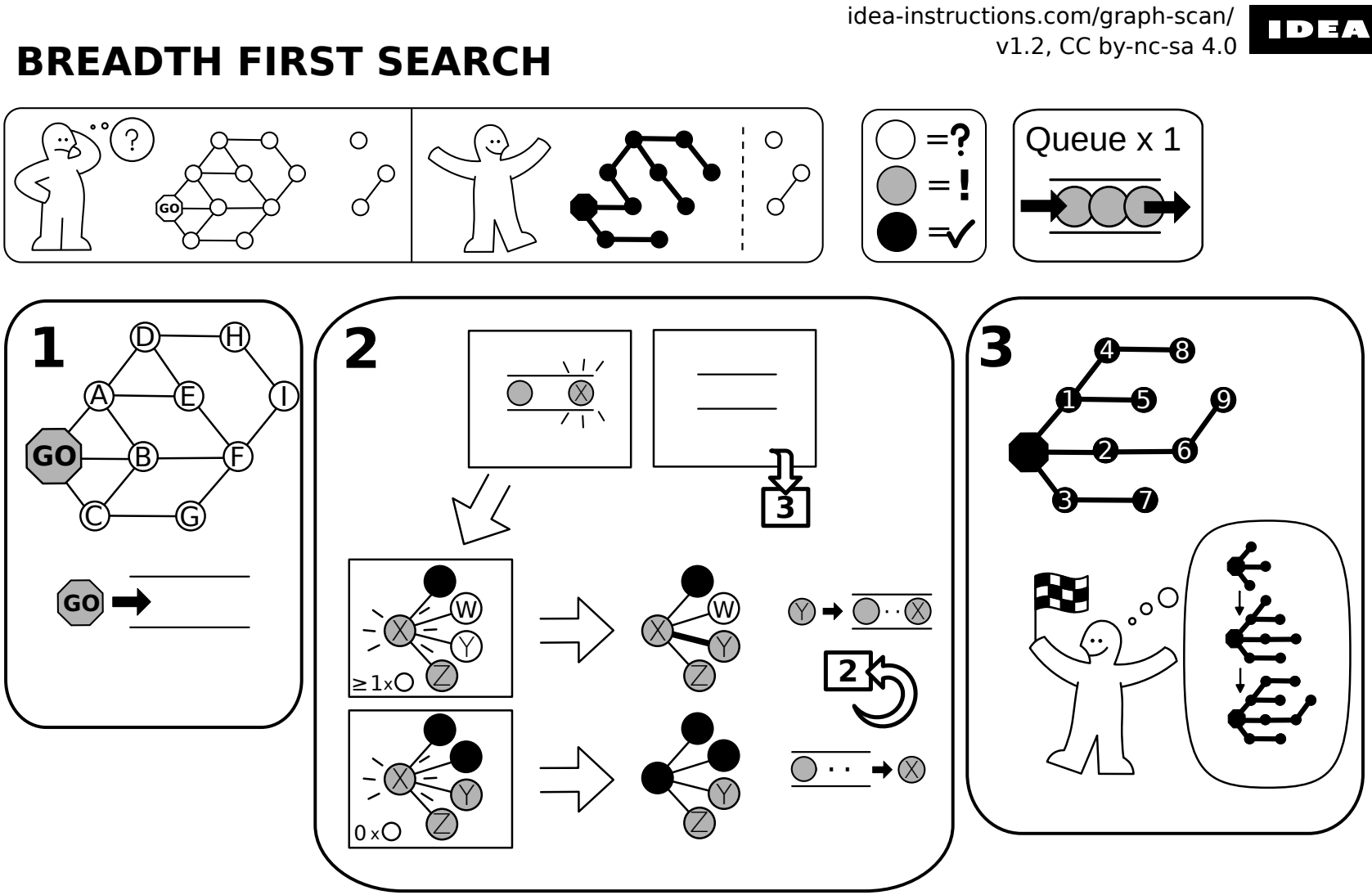
- La densité d'un graphe est le rapport:

$$\frac{2 \times m}{n \times (n - 1)}$$

Mesure: Distance

- La distance entre 2 sommets est la longueur du *plus court* chemin les reliant (plus d'un chemin de cette longueur possible)
- S'il n'y a *pas de chemin* alors la distance est considèrée comme *infinie*
- L'algorithme du parcours en largeur permet de calculer les distances d'un sommet x aux autres sommets

Algorithme: Parcours en largeur



Mesure: Eccentricité

- L'eccentricité d'un sommet est la plus grande distance qui le sépare des autres sommets.
- Le diamètre d'un graphe correspond à la plus grande valeur d'eccentricité que peut prendre un sommet du graphe (=la distance entre les 2 sommets les plus éloignés).
- Le rayon d'un graphe correspond à la plus petite valeur d'eccentricité que peut prendre un sommet du graphe
- La périphérie du graphe = sommets dont l'eccentricité est égale au diamètre
- Le centre du graphe = sommets dont l'eccentricité est égale au rayon

Mesure: Centralité de proximité

Closeness centrality

- C'est une mesure qui caractérise un noeud x d'un graphe $G = (V, E)$
- Soit n: nombre de noeuds dans le graphe

$$C(x) = \frac{n-1}{\sum_{y \in V} d(x,y)}$$

Intuition

$C(x) = 1/$ La distance moyenne de x aux autres noeuds du graphe.
Quand distance moyenne $\searrow$ centralité $\nearrow$.

Mesure: Centralité d'intermédiarité

Betweeness centrality:

- C'est une mesure qui caractérise un noeud x d'un graphe $G = (V, E)$

$$g(x) = \sum_{y \neq x \neq z} \frac{\sigma_{yz}(x)}{\sigma_{yz}}$$

où σ_{yz} : nombre de plus court chemins entre y et z, $\sigma_{yz}(x)$: nombre de plus court chemins entre y et z qui passent par x.

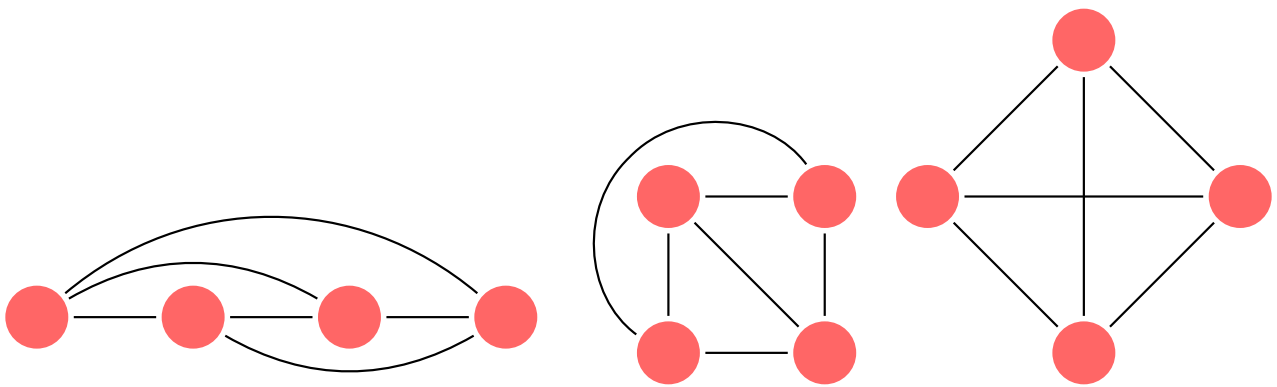
Intuition:

$g(x) \approx$ proportion de chemins reliant 2 sommets qui passent par x.
Dans le graphe des chemins de fers français Paris est central car le plus court chemin de n'importe quelle ville à une autre passe presque toujours par Paris.

Représentation

Un graphe peut avoir plusieurs représentations spatiales: Rien dans la définition d'un graphe n'indique quelles devraient être ses coordonnées dans un repère cartésien

Différents graphes?

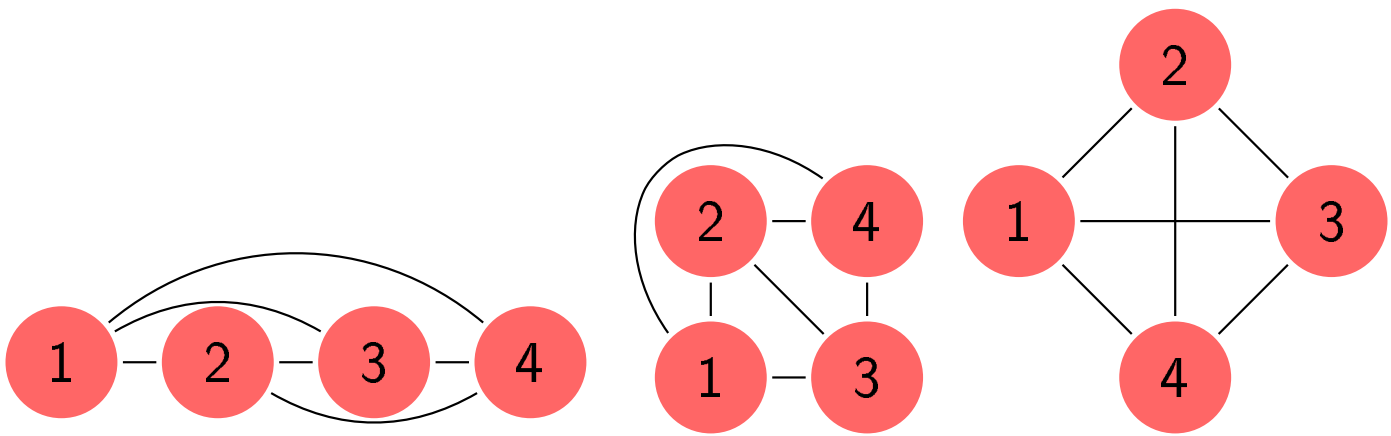


Représentation

Un graphe peut avoir plusieurs représentations spatiales: Rien dans la définition d'un graphe n'indique quelles devraient être ses coordonnées dans un repère cartésien

Un seul graphe! Plusieurs représentations.

$V = \{1, 2, 3, 4\}$
$E = \{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\} \}$



Représentation

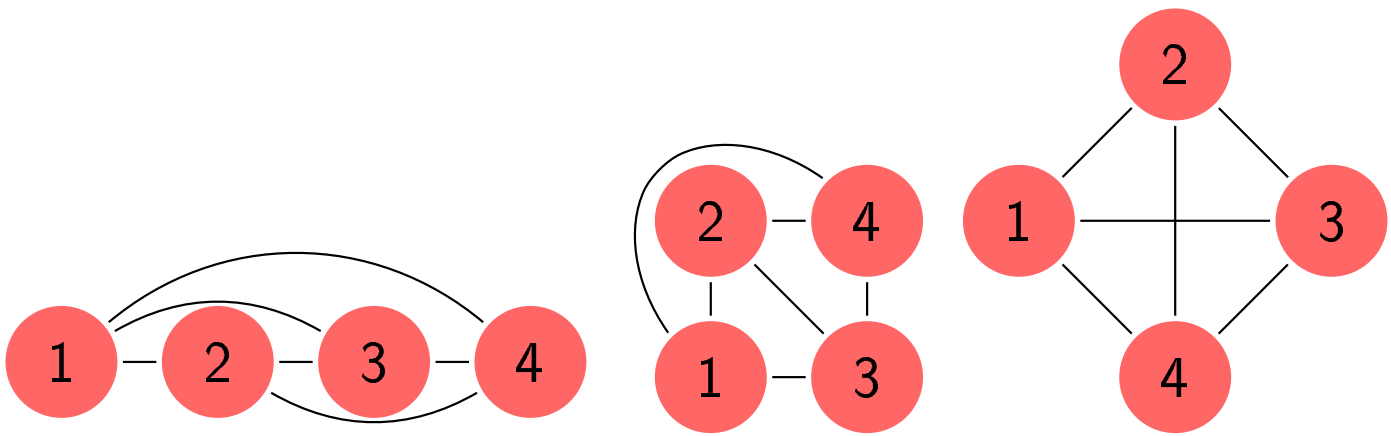
Un graphe peut avoir plusieurs représentations spatiales:

- Plusieurs positions possibles pour les noeuds
- Des arêtes qui se croisent ou non

Un seul graphe! Plusieurs représentations.

$V = \{1, 2, 3, 4\}$

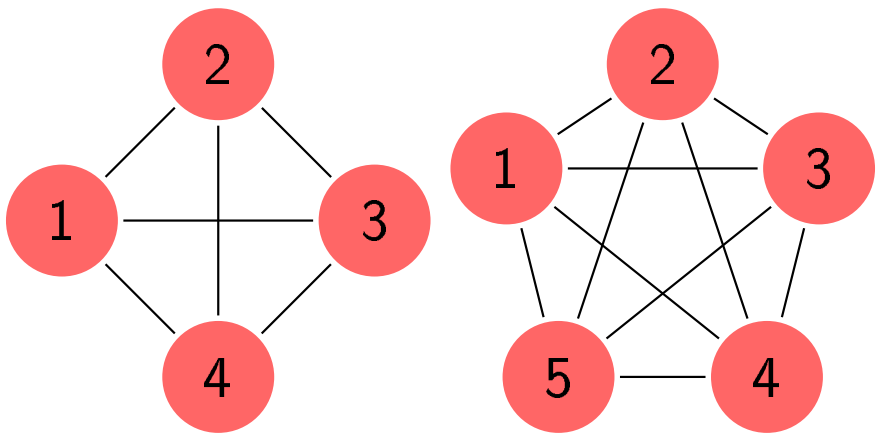
$E = \{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\} \}$



Représentation - planarité

S'il existe une représentation du graphe sans que les arêtes ne se croisent: Graphe planaire.

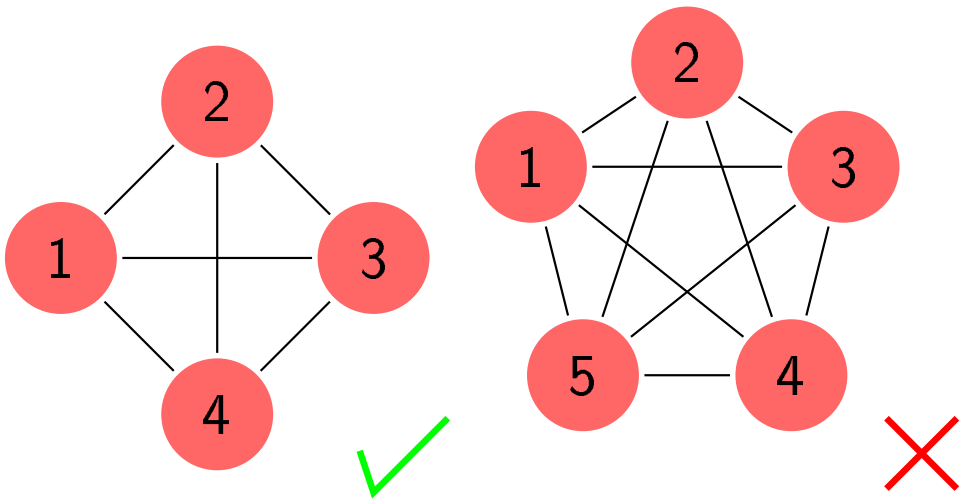
Est-ce que ces graphes sont planaires?



Représentation - planarité

S'il existe une représentation du graphe sans que les arêtes ne se croisent: Graphe planaire.

Est-ce que ces graphes sont planaires?

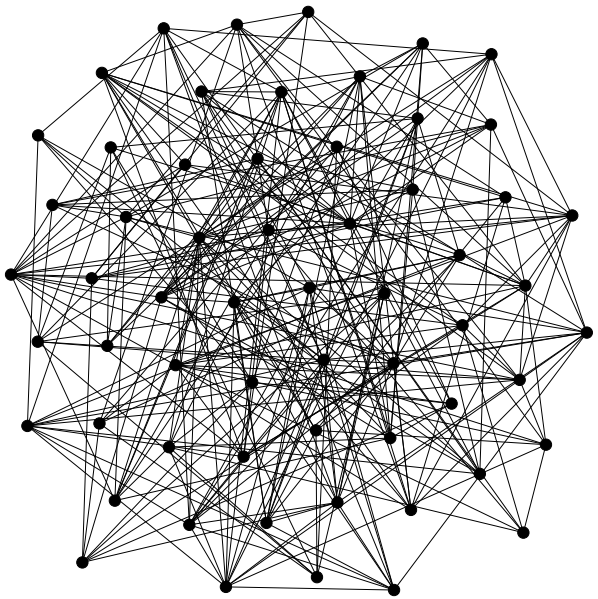


- Intuition: Plus un graphe est grand et dense, moins il y a de chances qu'il soit planaire

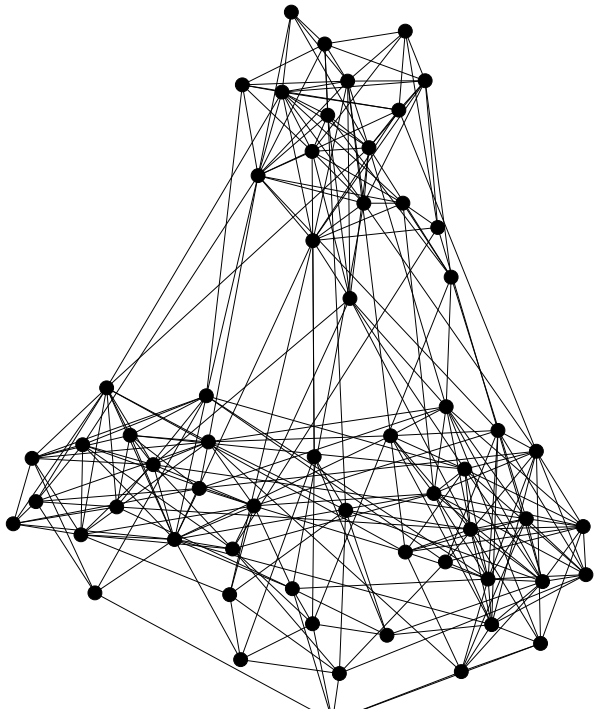
Agencement

Plusieurs algorithmes existent qui produisent un agencement (layout) du graphe dans le plan.
On se concentre pour cette présentation sur un algorithme implémenté dans Gephi:

au Hasard



ForceAtlas2



Agencement: Modèle d'energie

Force-directed layout

ForceAtlas2 est un algorithme d'agencement dirigé par la *simulation* d'un système physiques de forces qui s'exercent sur les sommets et les arêtes:

- attraction : les arêtes sont comme des ressorts ($F_a = -kd$)
- répulsion : les sommets non-adjacent se repoussent comme des particules chargées ($F_r = k/d^2$)
- Ces forces sont dépendantes des distances spatiales d entre les sommets: $d \searrow$, $F_a \searrow$, $F_r \nearrow$ et vice versa
- Les positions des sommets sont calculées et recalculées jusqu'à obtenir un équilibre: *des densités visuelles qui représentent les densités structurelles.*

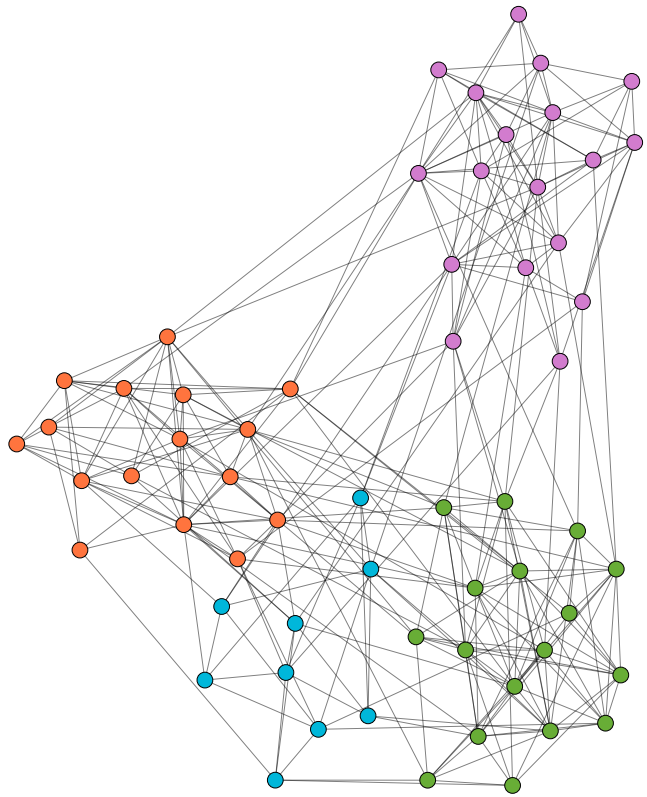
Partition d'un graphe

- Étant donné r catégories $C_1, \ldots, C_r$ la partition d'un graphe est l'assignation de chaque sommet du graphe à une des r catégories.
- L'union des r catégories couvre tous les sommets du graphe:

$$C_1 \cup \ldots \cup C_r = V$$

- Les catégories n'ont pas de sommets en commun.

$$C_i \cap C_j = \varnothing$$



Partition d'un graphe

Comment assigne-t-on une catégorie a un sommet?
Plusieurs cas:

- Le nombre de catégories est connu/inconnu: min cut, détection de communauté
- Les catégories attendues sont de taille fixée
- La catégorie de certains sommets est connue label propagation
- Hypothèse: le graphe a été généré par une distribution de probabilité

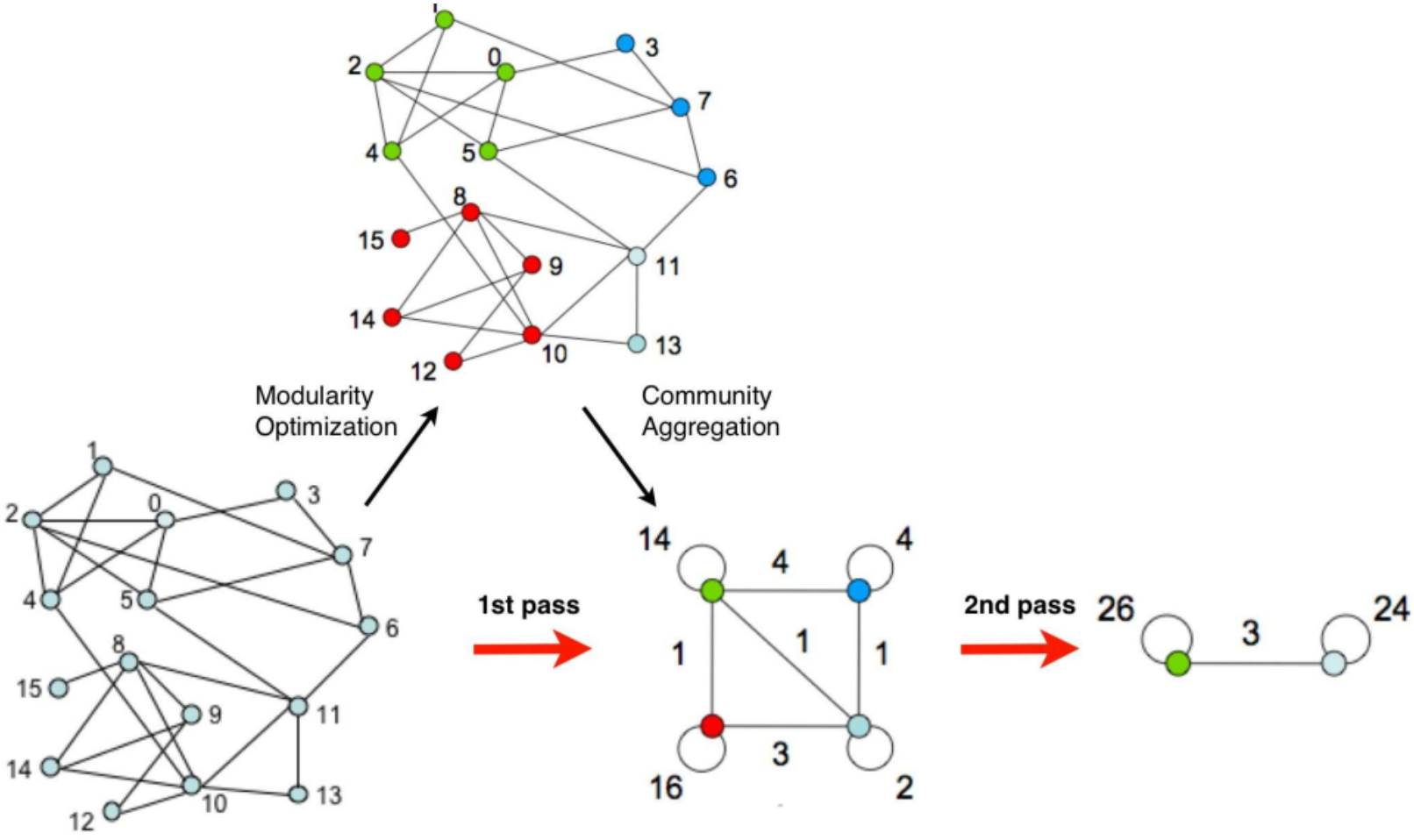
Modularité

- La modularité est une valeur entre -1 et 1 qui caractérise une partition de graphe
- Basée sur la densité des liens intra-catégorie et inter-catégories

$$Q = \frac{1}{2m} \sum_{x,y} (A_{xy} - \frac{d(x)d(y)}{2m}) \delta(c_x, c_y)$$

- c_x, c_y désignent les catégories des sommets x, y
- $\delta(c_x, c_y) = 1$ si les catégories sont les mêmes, 0 sinon
- L'algorithme de Louvain est un algorithme itératif qui retourne une partition de modularité maximale

Modularité



Source: Fast unfolding of communities in large networks. *Blondel et al.* Journal of Statistical Mechanics: Theory and Experiment. *2008.*

Conclusion

- Les graphes sont des objets mathématiques
- Les graphes sont caractérisables à l'aide de mesures qualitatives (eg graphe dirigé, complet, etc) ou quantitatives (eg calculer la centralité de ses sommets)
- Il existe diverses façons de visualiser des graphes. Les agencements "force-directed" des noeuds font ressortir les zones de densités structurelles du graphe
- Les algorithmes de détection de communautés permettent de partitionner le graphe en diverses zones de forte densité.
- Les graphes peuvent être conçus comme outils d'exploration des données
- Le choix de la sémantique de la relation est essentiel !
- Il existe beaucoup d'autres outils associés aux graphes notamment ceux modélisant des relations dynamiques