Big Data et Graphes : Quelques pistes de recherche

Hamamache Kheddouci

http://liris.cnrs.fr/hamamache.kheddouci

Laboratoire d'InfoRmatique en Image et Systèmes d'information

LIRIS UMR 5205 CNRS/INSA de Lyon/Université Claude Bernard Lyon 1/Université Lumière Lyon 2/Ecole Centrale de Lyon

http://liris.cnrs.fr

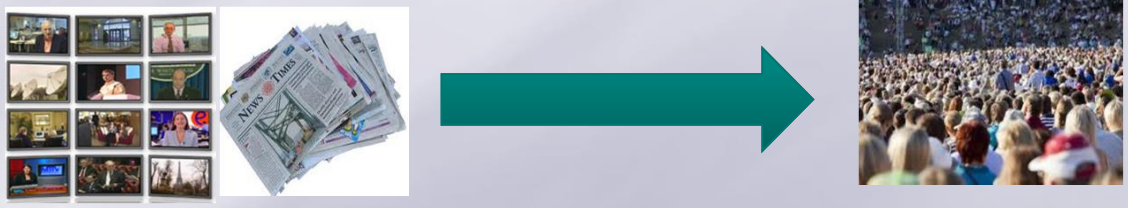
Big Data : Grandes Masses de Données

Age du Big Data !

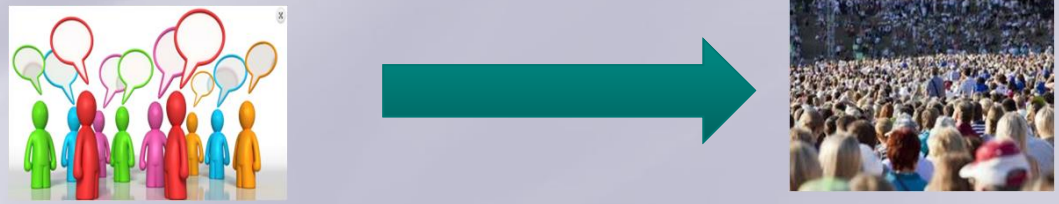
Nouveau modèle de données

Le Modèle de Génération/Consommation de la donnée a changé

Ancien modèle : Quelques compagnies génèrent des données, les autres sont des consommateurs de données



Nouveau Modèle : nous sommes tous des générateurs de données, et nous sommes tous des consommateurs de données



Générateurs des Big Data

Instruments scientifiques
(collecter toute sorte de données)

Mobiles
(tracer tous les objets tout le temps)

Média et réseaux sociaux
(tous des générateurs de données)

Réseaux de capteurs
(mesurer tout type de données)

Générateurs des Big Data

Instruments scientifiques
(collecter toute sorte de données)

Mobiles
(tracer tous les objets tout le temps)

Média et réseaux sociaux
(tous des générateurs de données)

Réseaux de capteurs
(mesurer tout type de données)

Age du Big Data

ñData is a new class of economic asset, like currency and gold.ò

Source: World Economic Forum 2012

Big Data

- **Un enjeu scientifique important :**

Big Data

Définitions é

ñ***Big Data***ò is a massive volume of both structured and unstructured data that is so large that it's difficult to process with traditional database and software techniques.ò

ñ***Big Data***ò is data whose scale, diversity, and complexity require new architectures, models, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from ité

Avec quels modèles ?

Big Data & Graphs ?

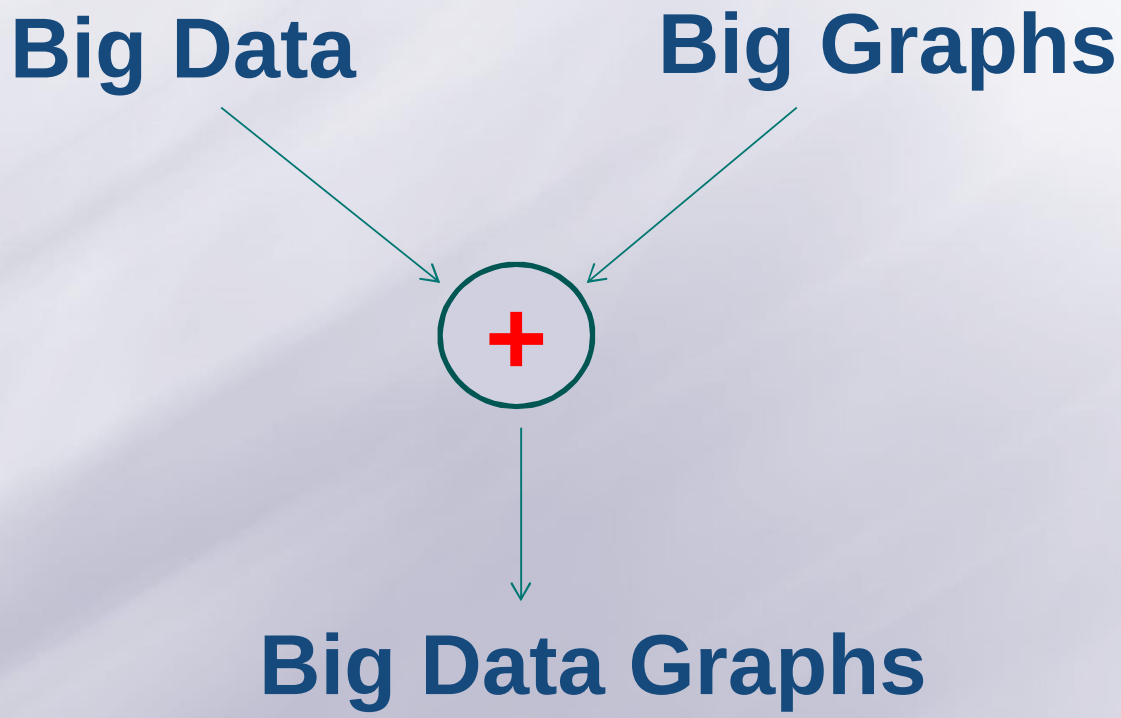
Naturellement, les graphes et les données sont liés :

- **Linked open Data (graphe dônteraction entre données)**
- **Des objets du Web sont des graphes (XML, RDF, é)**
- **Graphes des amis de Facebook**
- **Graphe de connaissances de Google**
- **Graphes extraits de grandes base de données**

Base de données

Données	temps	Emetteur	Récepteur	Type de mess.	….	Attribut n.
D1	1	S1	S2	A	….	3444
D2	2	S1	S3	C	….	2112
D3	3	S2	S4	B	….	5858
D4	4	S4	S2	A	….	600
D5	5	S3	S5	C	….	2333
….	….	….	….	….	….	….

Big Data & Big Graphs



Big Graphs for Big Data

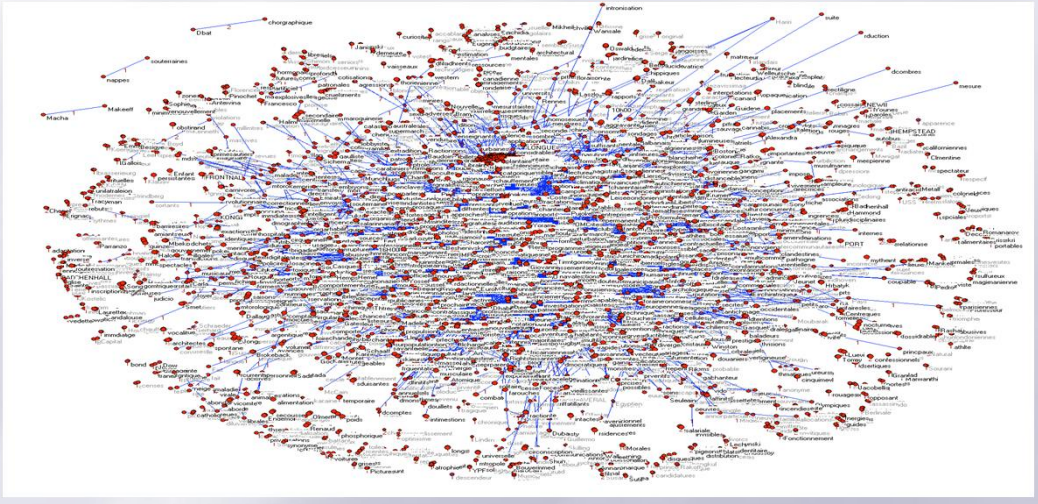
Verrous Big Data	Solutions à base de graphes
Indexation et stockage	Partitionnement de graphes
Flux de données (Vélocité)	Analyse de flux de graphes
Visualisation des données	Visualisation de graphes

Big Graphs for Big Data

Verrous Big Data	Solution à base de graphes
Indexation et stockage	Partitionnement de graphes
Analyse de flux de données	Analyse de flux de graphes
Visualisation des données	Visualisation de graphes

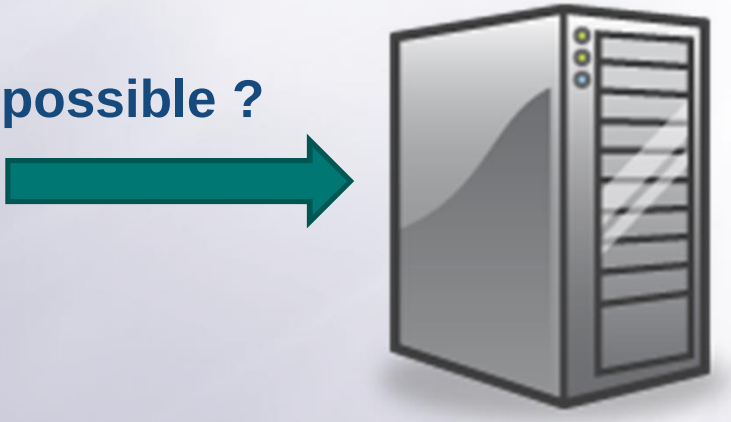
Big Graphs for Big Data

Partitionnement de graphes de données



Big Graphs
(Milliards de nî uds et arêtes)

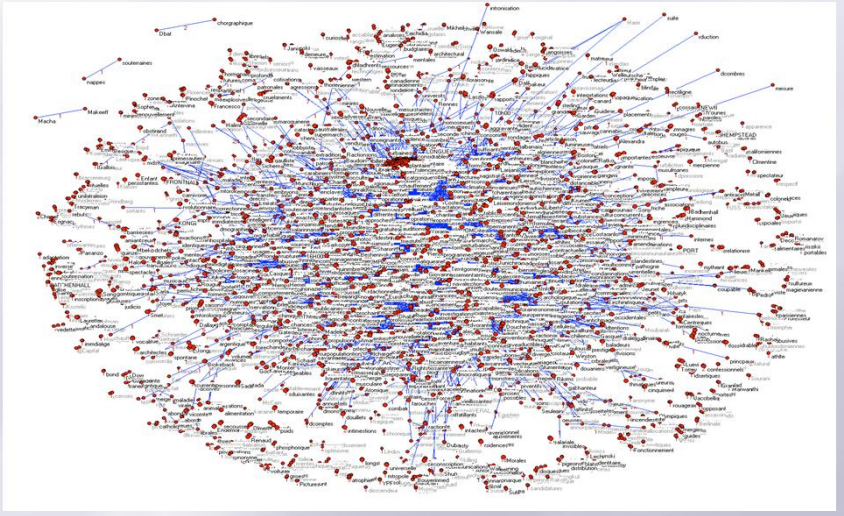
possible ?



1 machine ?
Combien de :
- RAM?
- Disque ?

Big Graphs for Big Data

- **Partitionnement de graphes de données**



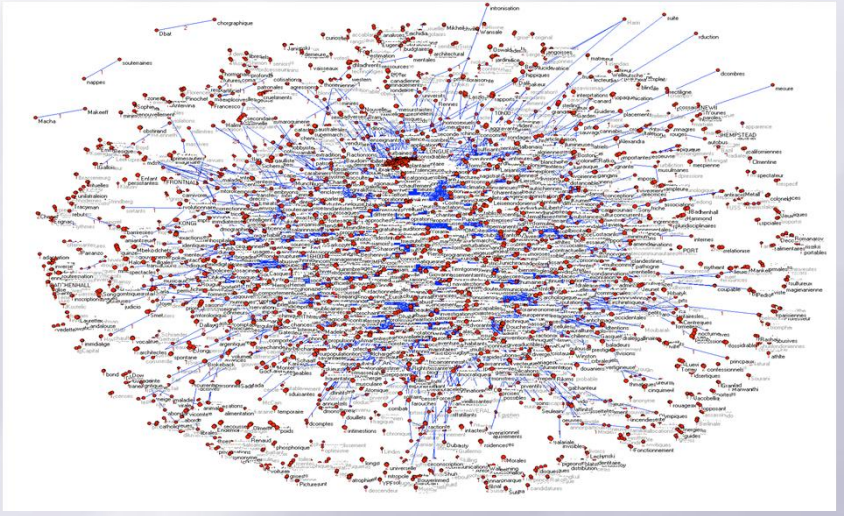
Big Graphs
(Milliards de nî uds et arêtes)

possible ?

k machines

Big Graphs for Big Data

- **Partitionnement de graphes de données**



Big Graphs
(Milliards de nî uds et arêtes)

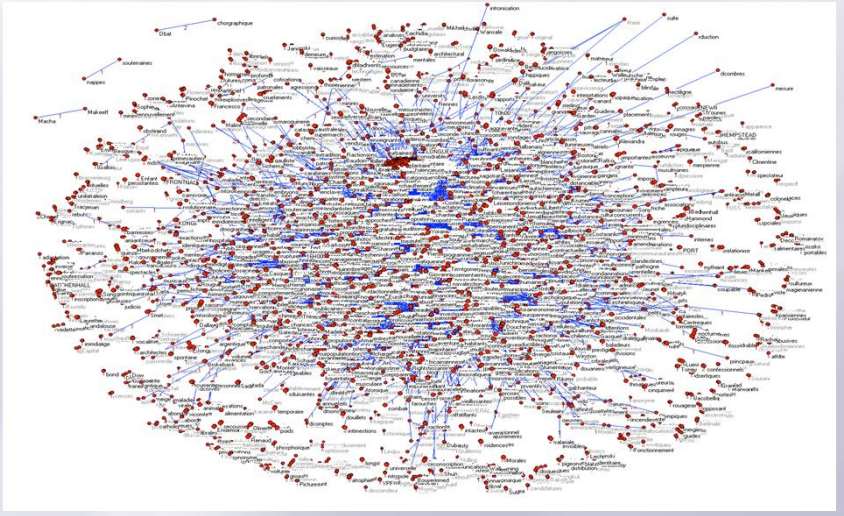
possible ?

OUI

k machines

Big Graphs for Big Data

Partitionnement de graphes de données



possible ?

OUI

Big Graphs
(Milliards de nî uds et arêtes)

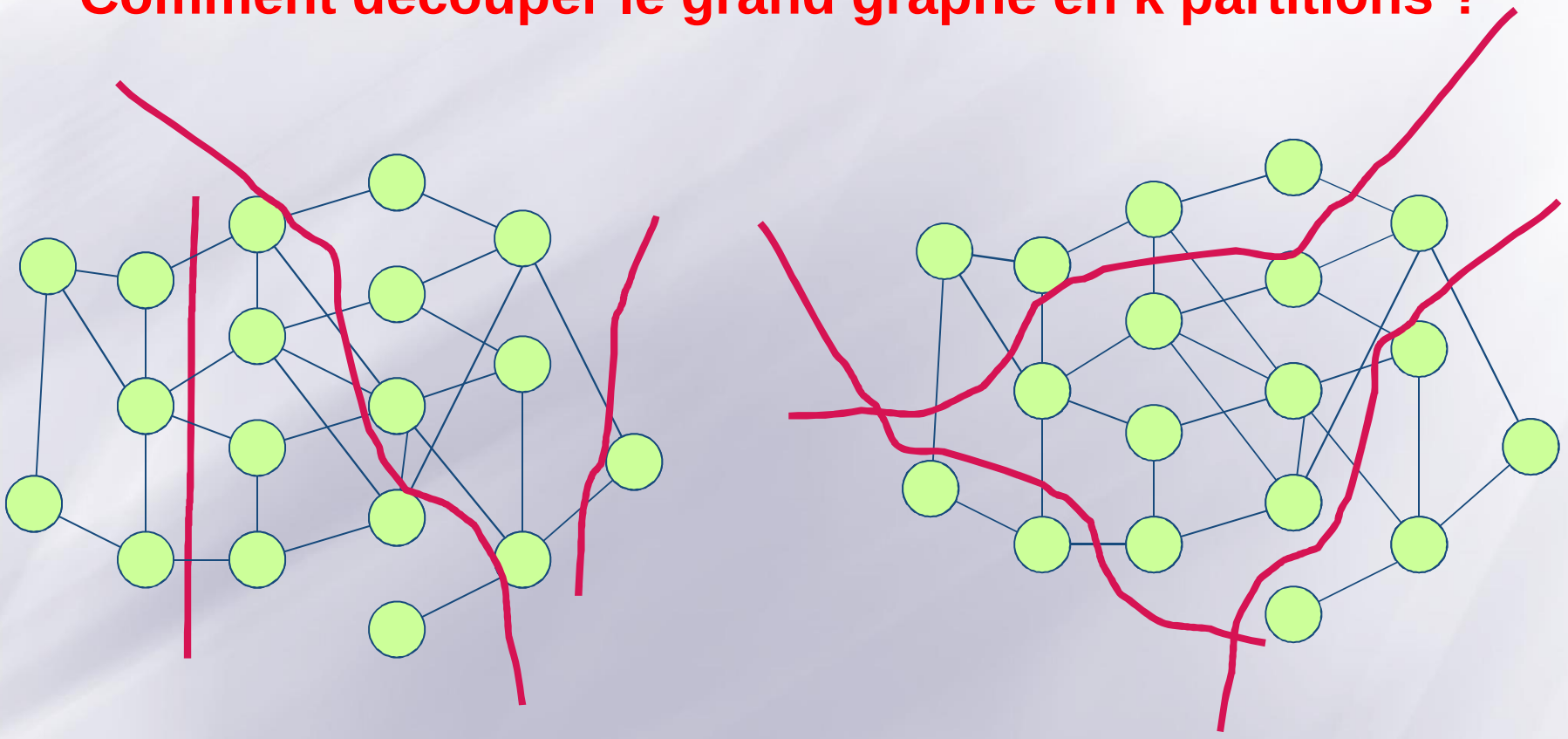
k clusters

Comment découper le grand graphe ?

Big Graphs for Big Data

- **Partitionnement de graphes de données**

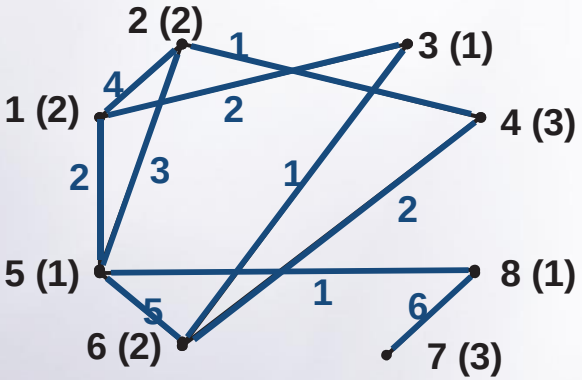
Comment découper le grand graphe en k partitions ?



Big Graphs for Big Data

Partitionnement de graphes de données

- Etant donné un graphe G = (N, E, W_N, W_E)
  - N = sommets,
  - W_N = poids sur les sommets
  - E = arêtes
  - W_E = poids des arêtes

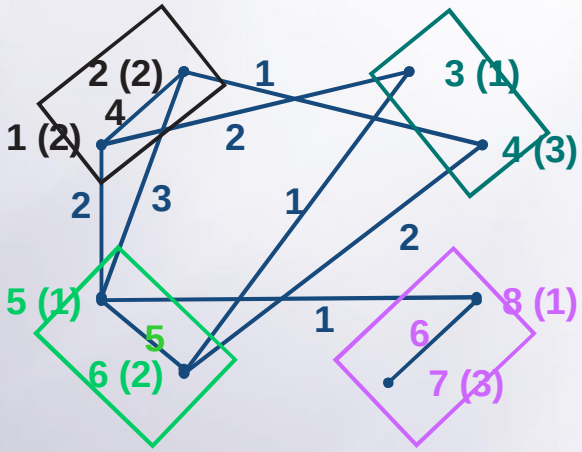


- Ex: N = {données}, W_N = {vecteurs d'attributs de données}, arête (j,k) dans E : j envoie W_E(j,k) mots au k
- Choisir une partition N = N_1 U N_2 U é U N_P telle que
  - La somme des poids des nœuds dans chaque N_j est à peu près le même
  - La somme des poids des arêtes connectant toutes les différentes paires N_j et N_k est minimisée
- Ex: équilibrage des chargements de données, en minimisant la communication entre les machines
- Cas particulier, N = N_1 U N_2

Big Graphs for Big Data

Partitionnement de graphes de données

- Etant donné un graphe G = (N, E, W_N, W_E)
  - N = sommets,
  - W_N = poids sur les sommets
  - E = arêtes
  - W_E = poids des arêtes



- Ex: N = {données}, W_N = {vecteurs d'attributs de données}, arête (j,k) dans E : j envoie W_E(j,k) mots au k
- Choisir une partition $N = N_1 \cup N_2 \cup \dots \cup N_P$ telle que
  - La somme des poids des nœuds dans chaque N_j est presque la même
  - La somme des poids des arêtes connectant toutes les différentes paires N_j et N_k est minimisée
- Ex: équilibrage des chargements de données, en minimisant la communication entre les machines
- Cas particulier, $N = N_1 \cup N_2$

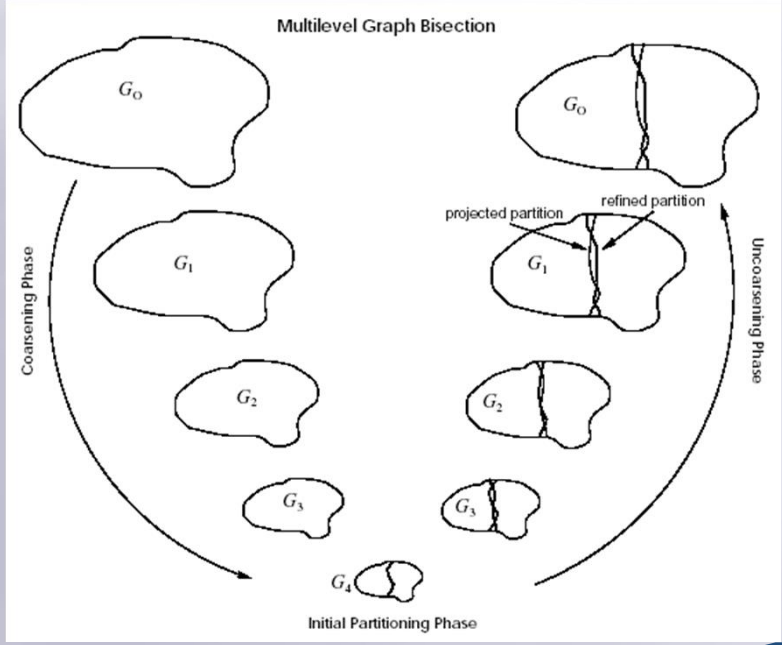
Big Graphs for Big Data

- **Partitionnement de graphes de données**
  - **NP-complet**
  - **Plusieurs algorithmes existent :**
    - **partitionnement spectral**
    - **partitionnement géométrique**
    - **partitionnement en graphes Multi-niveaux**

Big Graphs for Big Data

Partitionnement de graphes de données

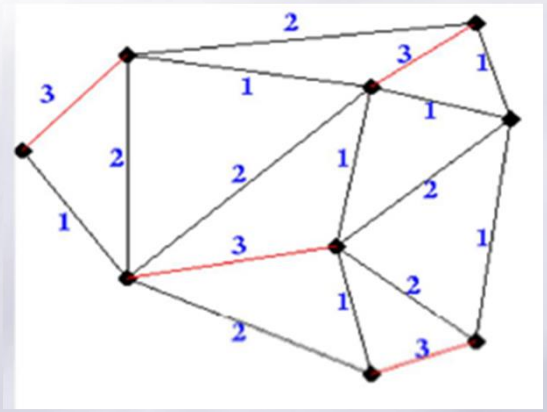
- NP-complet
- Plusieurs algorithmes existent :
  - partitionnement spectral
  - partitionnement géométrique
  - partitionnement en graphes Multi-niveaux
    - 3 Phases
      - compresser
      - Partitionner
      - Décompresser



Big Graphs for Big Data

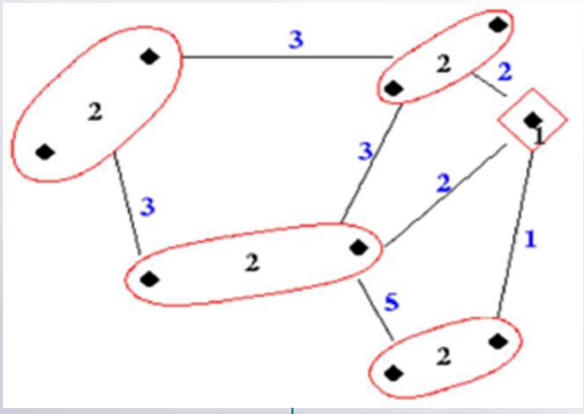
Partitionnement de graphes de données

Partitionnement en graphes Multi-niveaux

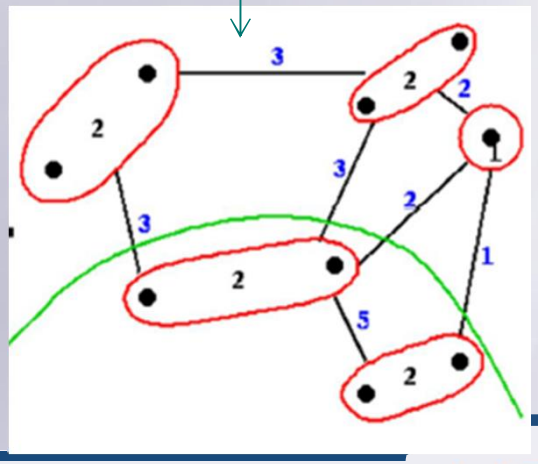


Compresser

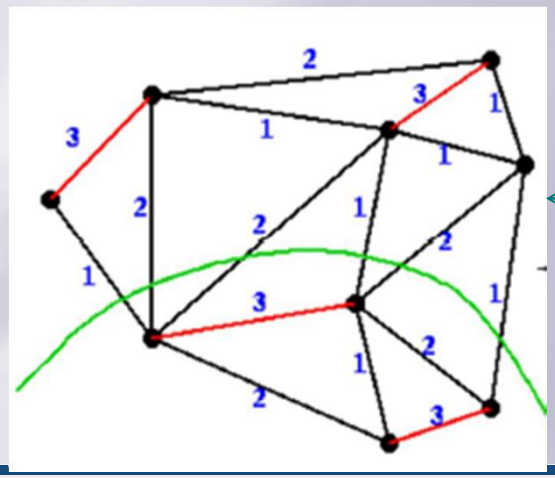
(couplage maximum)



partitionnement équilibré



décompresser



Big Graphs for Big Data

Partitionnement de graphes de données

Partitionnement à base du Lemme de régularité (Gastineau *et Kheddouci* 2014)

Lemme de régularité de Szemerédi

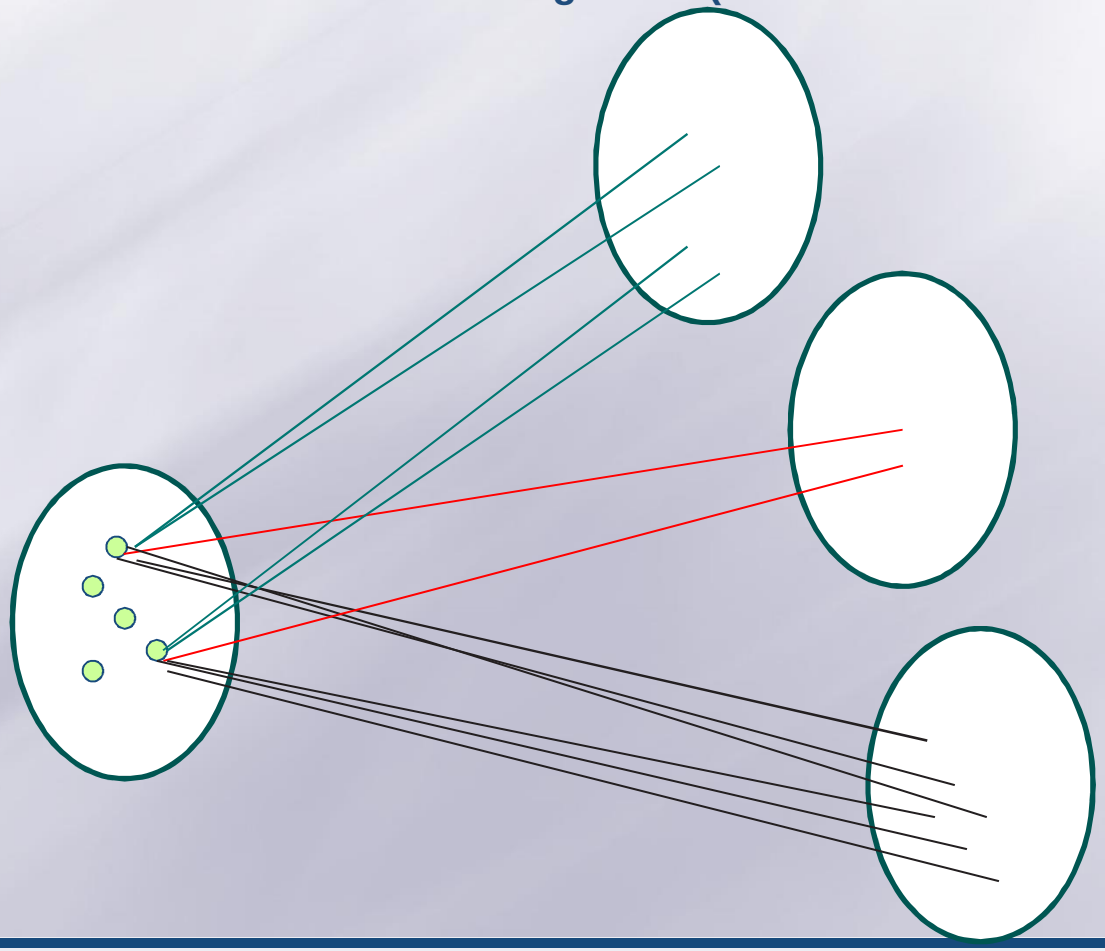
Pour tout réel positive $\epsilon > 0$ et tout entier positif t, il existe un entier $T = T(\epsilon, t)$ tel que tout graphe avec $n > T$ sommets à une partition en $k + 1$ classe telle que:

- $t \leq k$,
- $|V_i| < \lceil \epsilon n \rceil$,
- $||V_i| - |V_j|| \leq 1$ pour tous i,j,
- Tous les couples (V_i, V_j) sauf ϵk^2 sont ϵ-réguliers.

Big Graphs for Big Data

Partitionnement de graphes de données

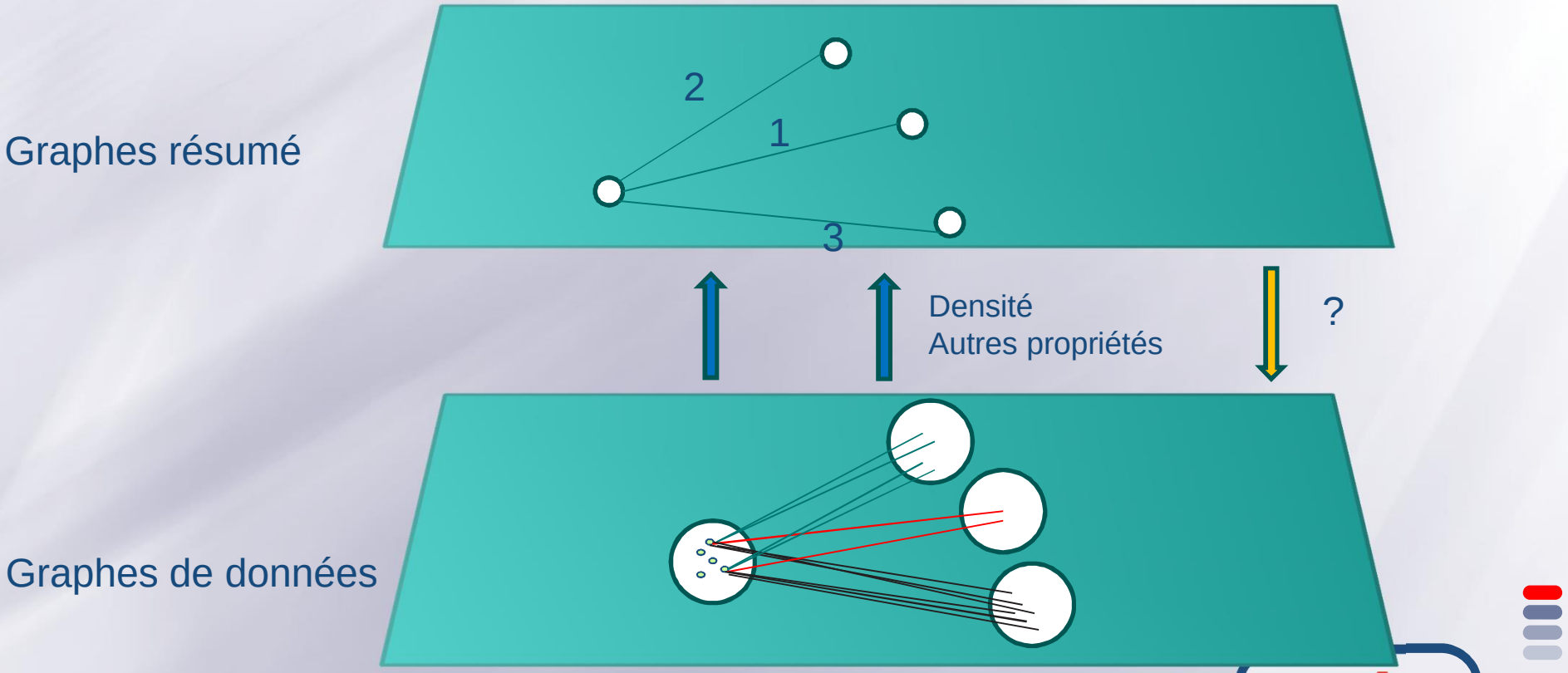
Partitionnement à base du Lemme de régularité (Gastineau *et Kheddouci* 2014)



Big Graphs for Big Data

Partitionnement de graphes de données

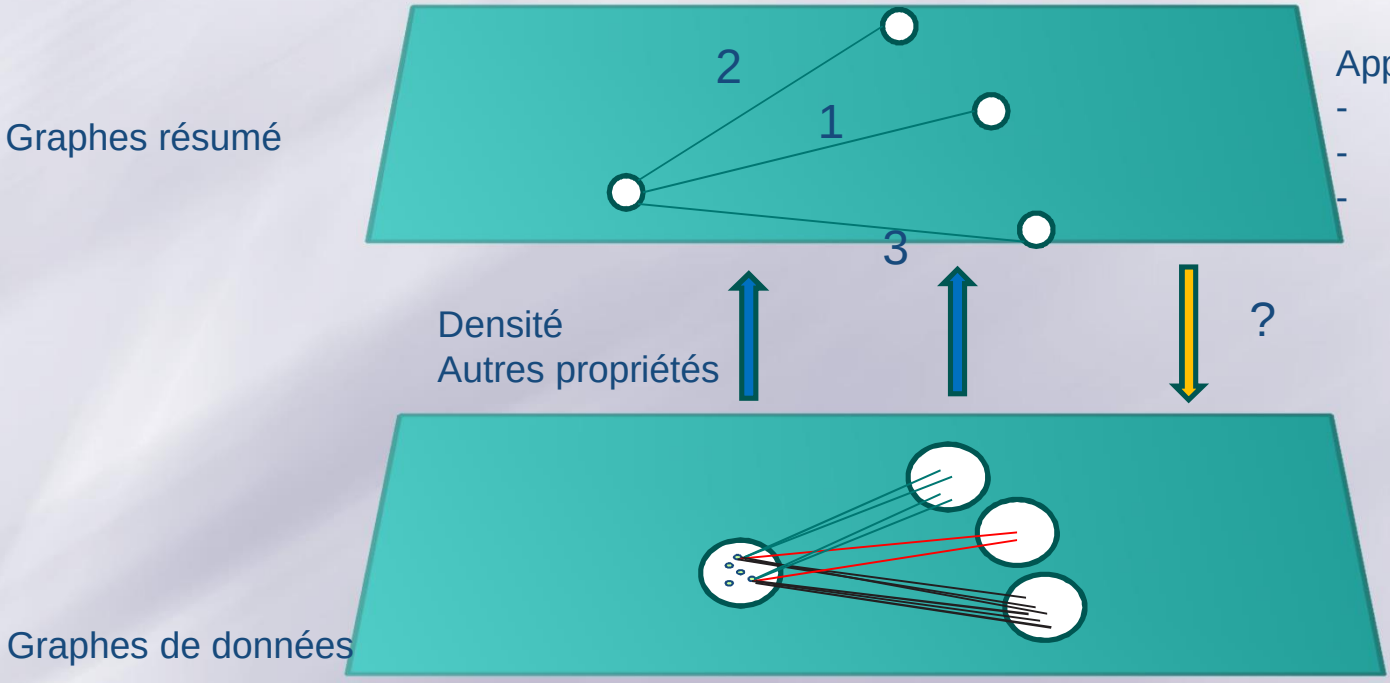
Partitionnement à base du Lemme de régularité (Gastineau *et Kheddouci* 2014)



Big Graphs for Big Data

Partitionnement de graphes de données

Partitionnement à base du Lemme de régularité (Gastineau *et Kheddouci* 2014)



Approximer :
- La recherche de motifs
- Construction de communautés
- é .

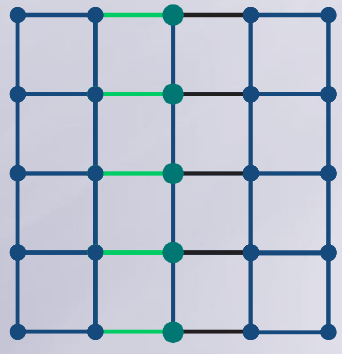
Big Graphs for Big Data

Partitionnement de graphes de données

Quelques paramètres de graphes liés au partitionnement

- **Ensemble d'arêtes Séparateur** : E_s (sous-ensemble de E) sépare G si en retirant E_s de E donne 2 composantes connexes de tailles égales, N: N_1 and N_2
- **Ensemble de sommets Séparateur** : N_s (sous-ensemble de N) sépare G si en retirant N_s et toutes leurs arêtes incidentes donne 2 composantes connexes de tailles égales, N: N_1 and N_2

G = (N, E), sommets N et arêtes E
N_s = sommets verts



Big Graphs for Big Data

Verrous Big Data	Solution à base de graphes
Indexation et stockage	Partitionnement de graphes
Analyse de flux de données	Analyse de flux de graphes
Visualisation des données	Visualisation de graphes

Big Graphs for Big Data

- Analyse de graphes de données en flux

Flux de données :

- Un flux de données est une séquence de données : $a_1, a_2, é , a_n$.
  - Flux de prix
  - Flux de paquets IP
- Les données ont différentes formes dans différentes applications.
  - Valeur scalaire
  - Tuple
  - é
- La sémantique des données est également différente dans différentes applications.

Big Graphs for Big Data

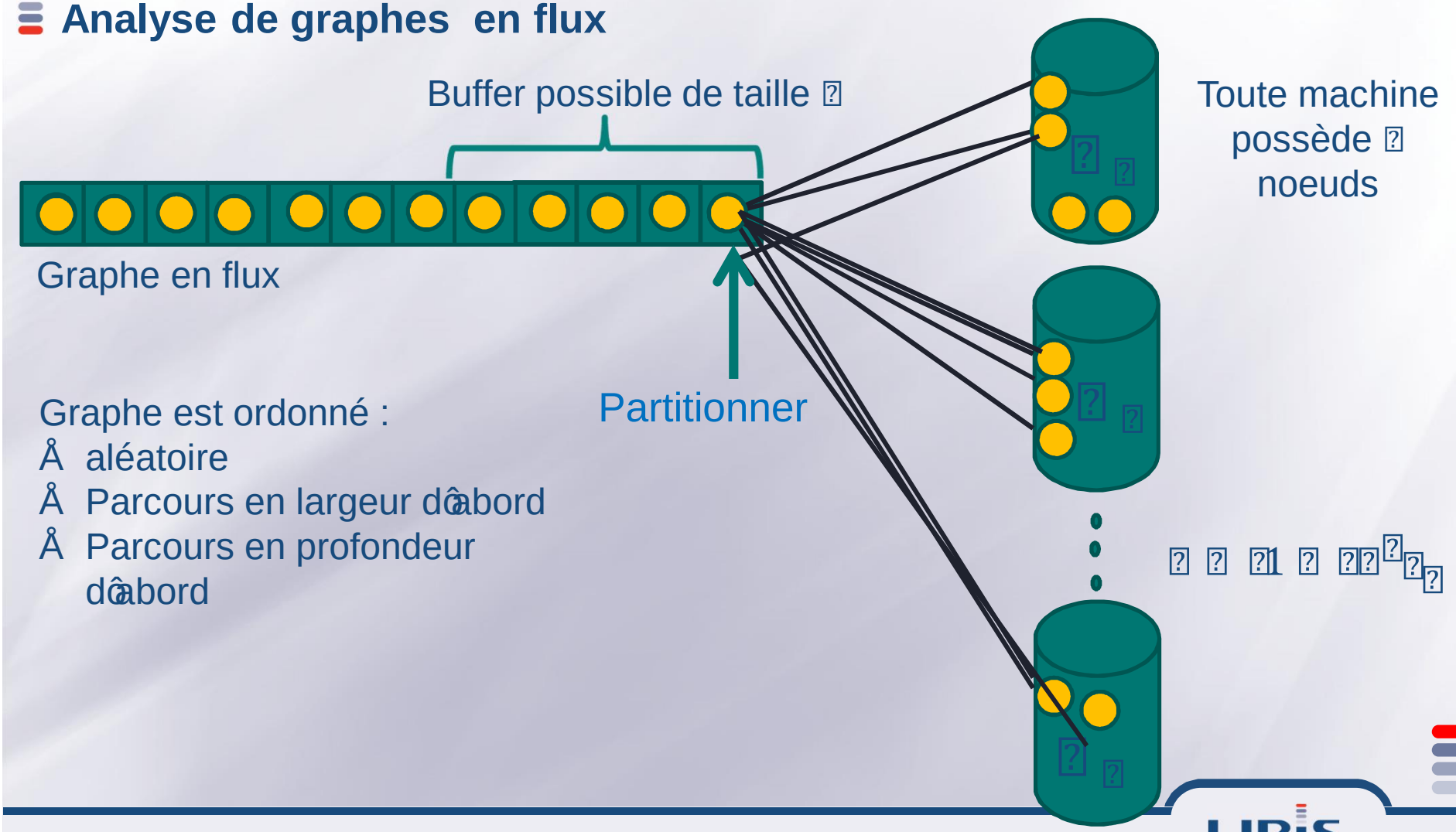
Analyse de graphes en flux

Modèle de traitement de flux :

- **Accès séquentiel** au flux de données
- **Ordre des données** dans le flux n'est pas contrôlé par l'algorithme et peut être artificiel.
- **Petit espace de travail** comparé à la longueur du flux n :
  - Polylog n
  - n^{ε}
- **Petit nombre de passes** sur le flux :
  - Une passe
  - Un nombre constant de passes
- **Temps de traitement d'une donnée est court**

Big Graphs for Big Data

- **Analyse de graphes en flux**



Graphe est ordonné :
- Å aléatoire
- Å Parcours en largeur dâabord
- Å Parcours en profondeur dâabord

Big Graphs for Big Data

- **Analyse de graphes en flux**
  - **Des travaux existent :**
    - **Tester la connectivité,**
    - **Tester la planarité,**
    - **construction dôarbre couvrant,**
    - **é**
  - **Re-penser les problèmes de graphes classiques dans le modèle streaming, notamment pour lôorganisation des données : clustering, détection de propriétés structurelles, etc**

Big Graphs for Big Data

- Analyse de graphes en flux
  - Heuristiques pour le partitionnement de graphes en flux :
    - Linear Deterministic Greedy LDG (Stanton and Kliot 2012)
    - FENNEL (Tsourakakis *et al.* 2013)
    - Restreaming graph partitioning (Nishimura *et al.* 2013)

Nouvelle Méthode : Partial Restreaming Model

Modèle du Restreaming partiel (Echbarthi et Kheddouci 2014)

Ce modèle consiste en 2 phases :

- **1ère phase: Restreaming dûne portion du graphe, dûne fenêtre de taille choisie.**
- **2ème phase: Le reste du graphe est traité en une seule passe du stream.**

Méthode proposée: Restreaming partiel

- **Dans le modèle du *restreaming partiel*, on sélectionne des portions du graphe de taille *n/k** qui vérifient des propriétés pertinentes dans le graphe :**

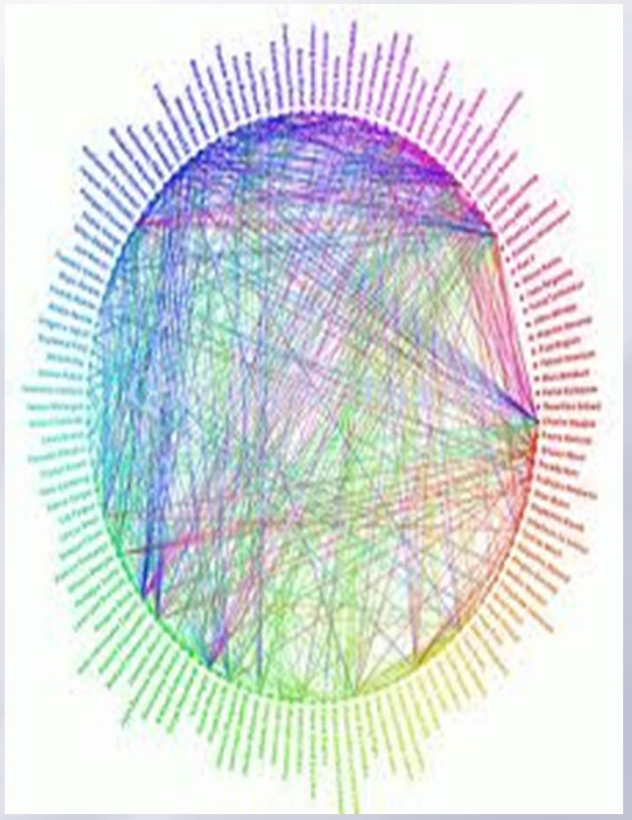
  - **Densité de la portion >= psi* densité du graphe**
  - **Degré moyen de la portion >= eta * degré moyen du graphe**
  - **é .**

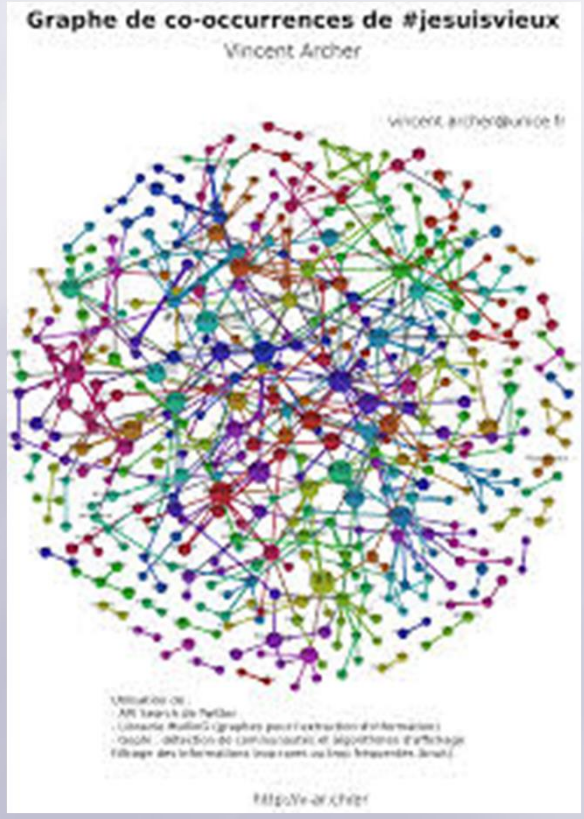
Big Graphs for Big Data

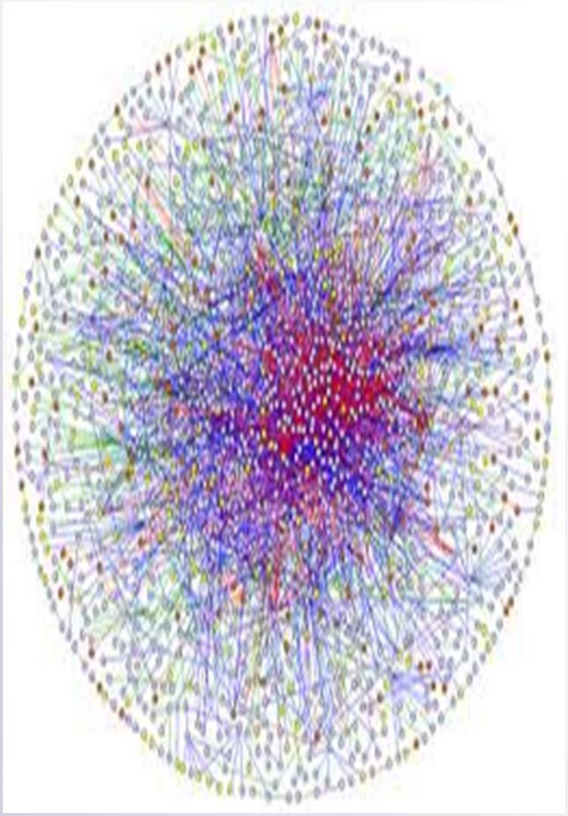
Verrous Big Data	Solution à base de graphes
Indexation et stockage	Partitionnement de graphes
Analyse de flux de données	Analyse de flux de graphes
Visualisation des données	Visualisation de graphes

Big Graphs for Big Data

Visualisation des grands graphes

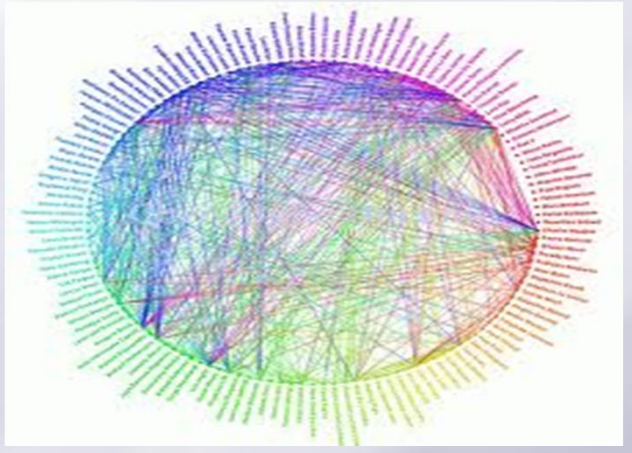


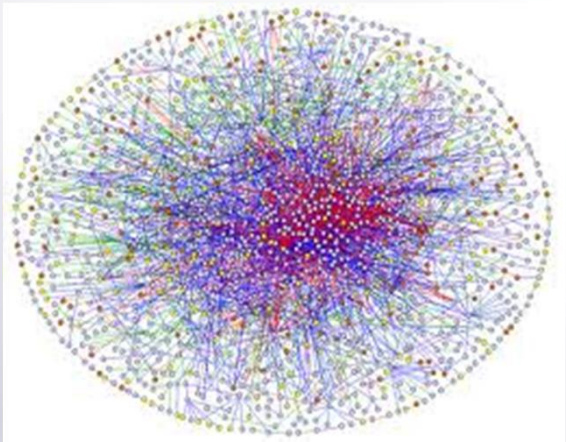




Big Graphs for Big Data

Visualisation de grands graphes



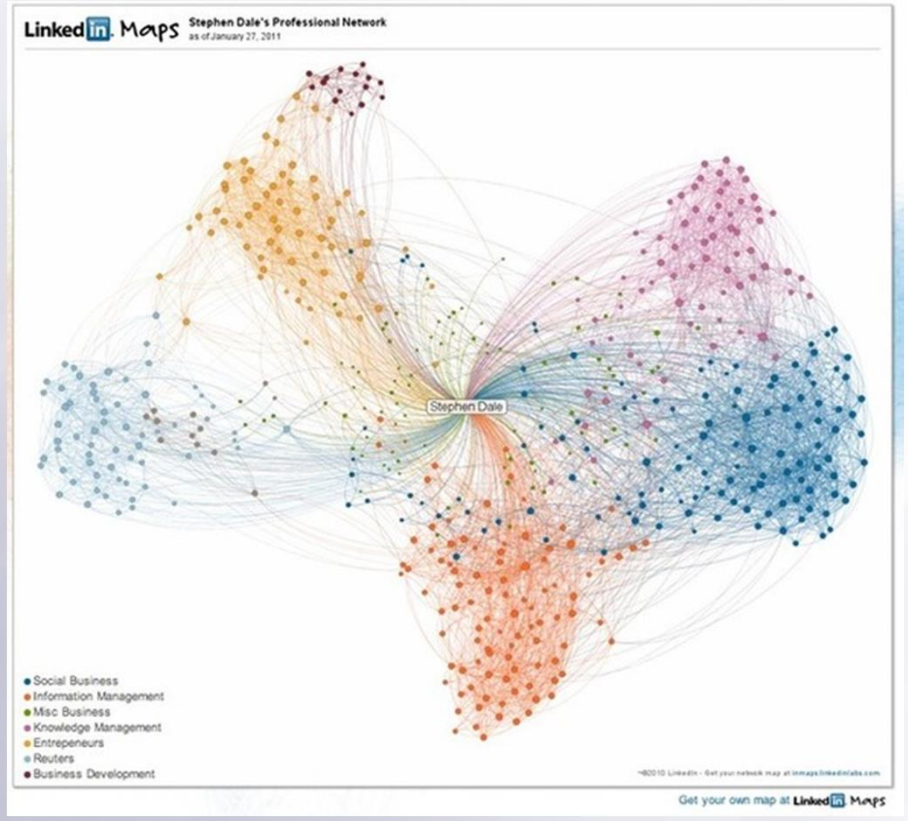


- Comment visualiser les grands graphes de données ?
- Quels algorithmes et techniques pour explorer visuellement le grand graphe ?

Big Graphs for Big Data

- **Visualisation de grands graphes**

- **Communautés/clustering/classification**
- **Recherche de motifs fréquents**
- **Visualisation d'échantillons représentatifs, de sous-graphes, etc**
- **Visualisation 2D, 3D**
- **Combiner la fouille visuelle avec l'exploration algorithmique des grands graphes.**



Big Data et Graphes,

Ce nôest que le début de lôhistoire é . !!