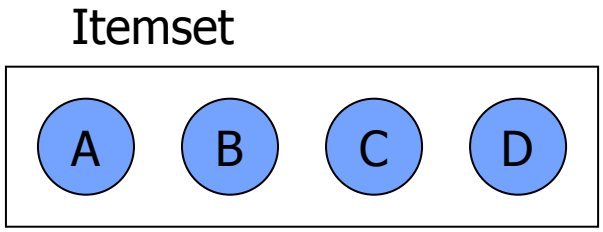
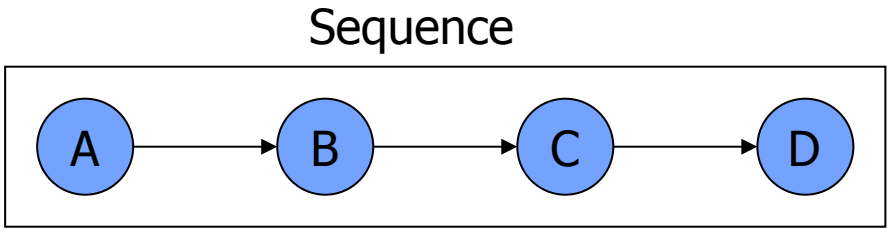
L I P 6 C N R S

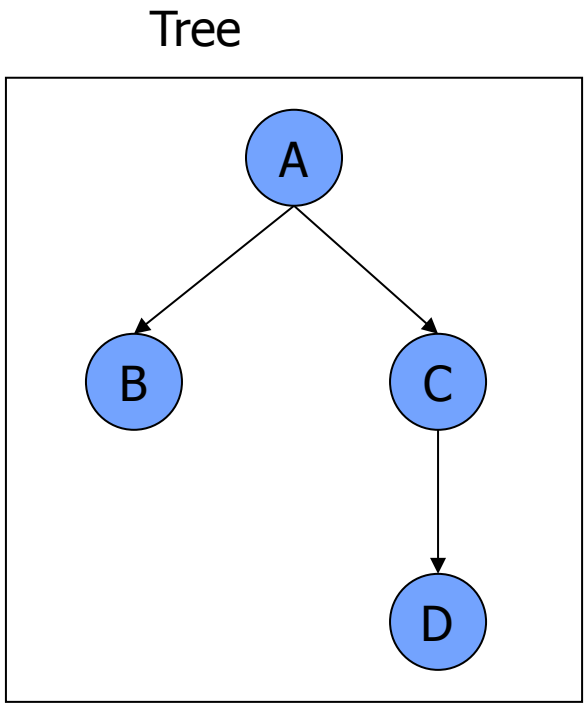
Fouille dans les graphes
Graph Mining

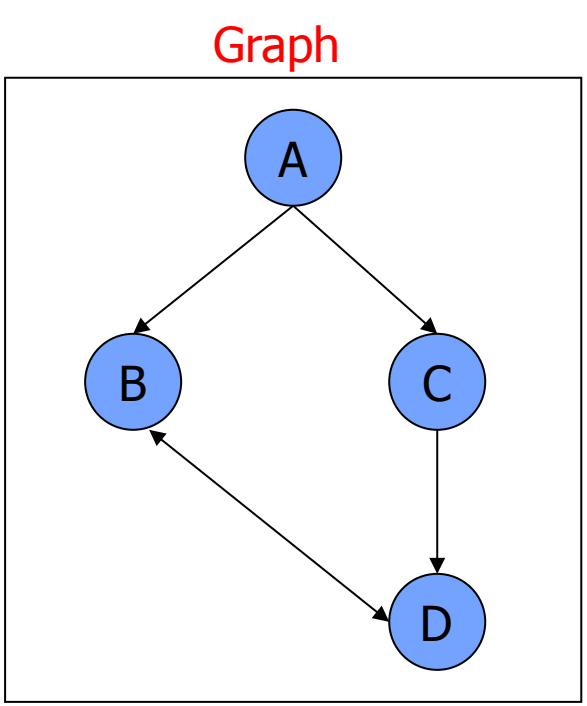
Caractérisation des grands graphes
Recherche de motifs fréquents

Example Pattern Types







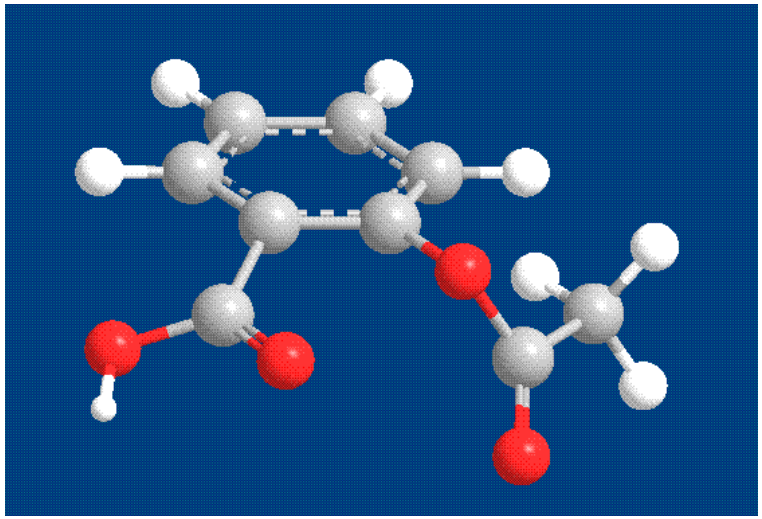


- Can add attributes
  - To nodes
  - To edges
- **Attributes**
  - Labels
  - Type (directed or undirected)
  - Set-valued

Fouille de graphes

- Problèmes
  - À quoi ressemblent les graphes
  - Comment évoluent-ils?
  - Quels outils utiliser?
  - Passage à l'échelle

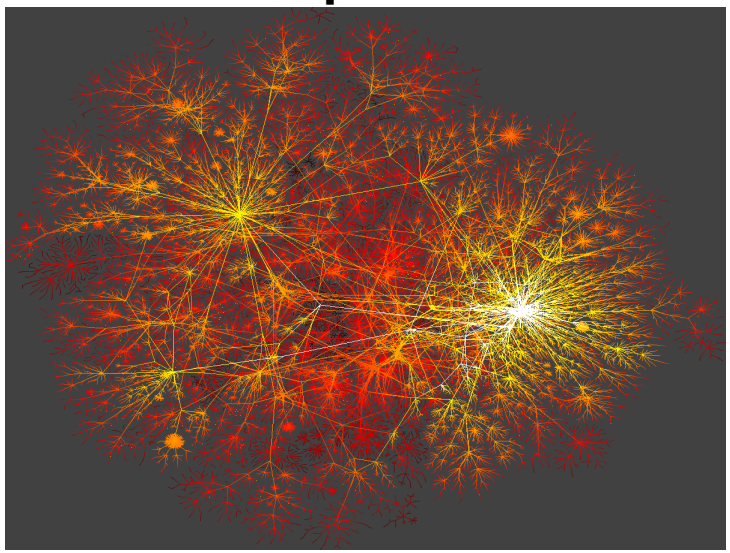
Présence des graphes…



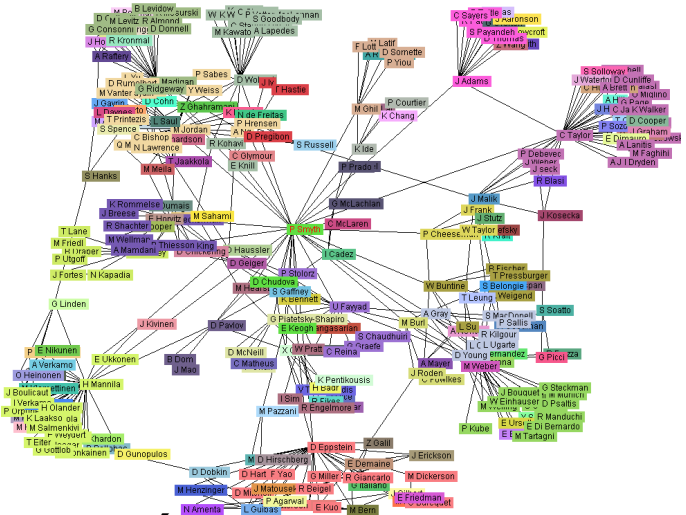
Aspirine

from H. Jeong et al Nature 411, 41 (2001)

réseau d'interaction de protéines



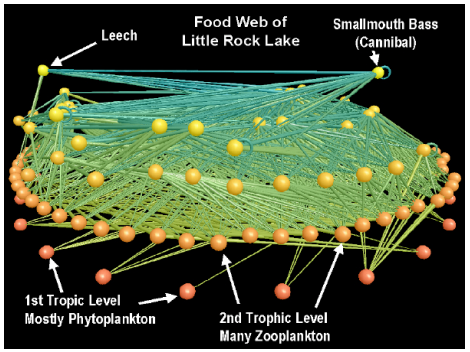
Internet



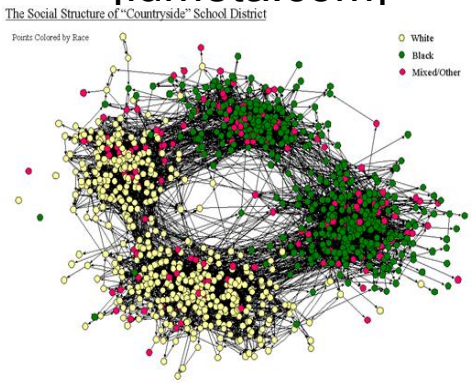
Réseau de co-auteurs

Graphes - Exemples

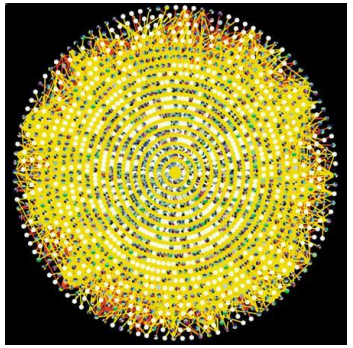
Internet Map
[lumeta.com]



Food Web
[Martinez ' 91]



Friendship Network
[Moody ' 01]

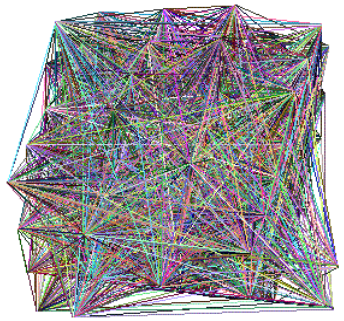


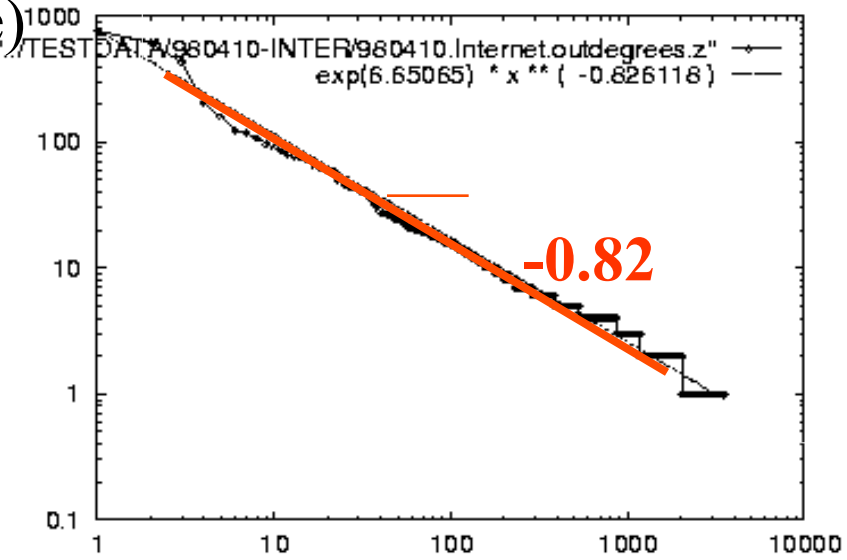
Protein Interactions
[genomebiology.com]

Propriétés
Distribution des degrés des noeuds

internet domains

log(degree)

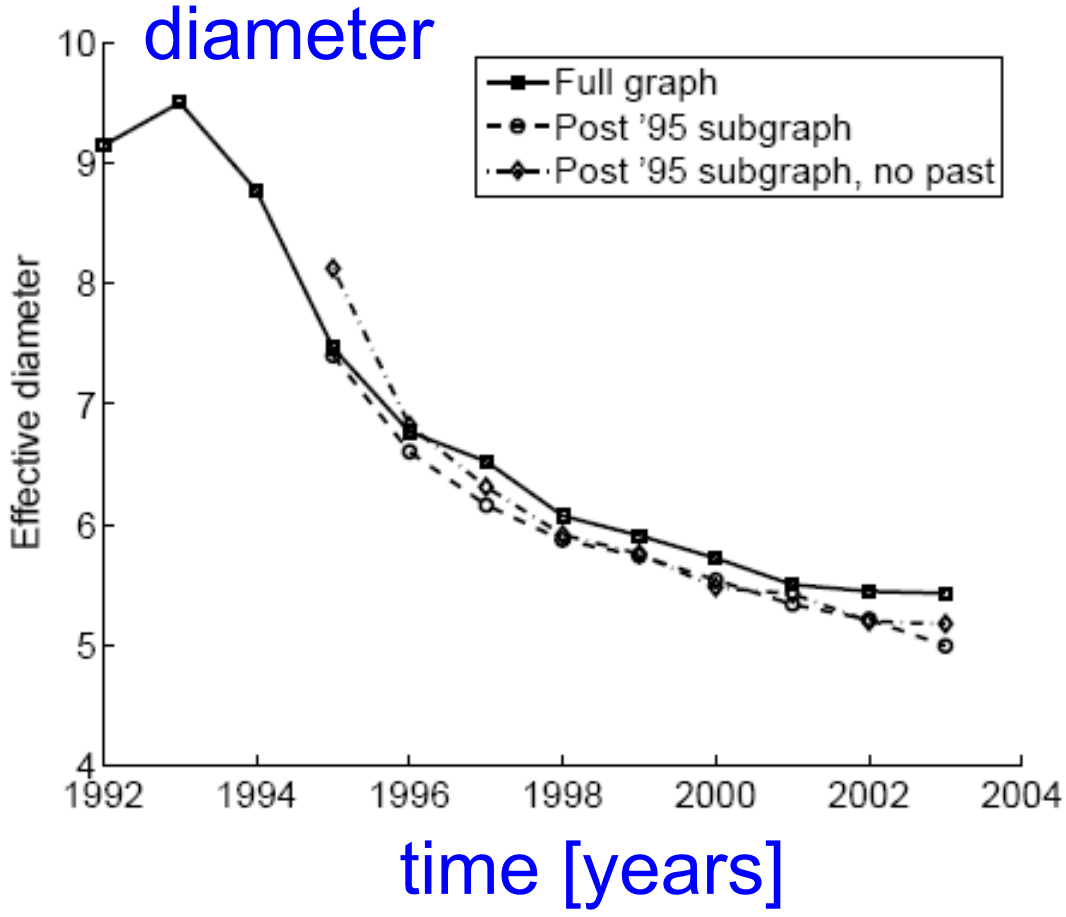




log(rank)

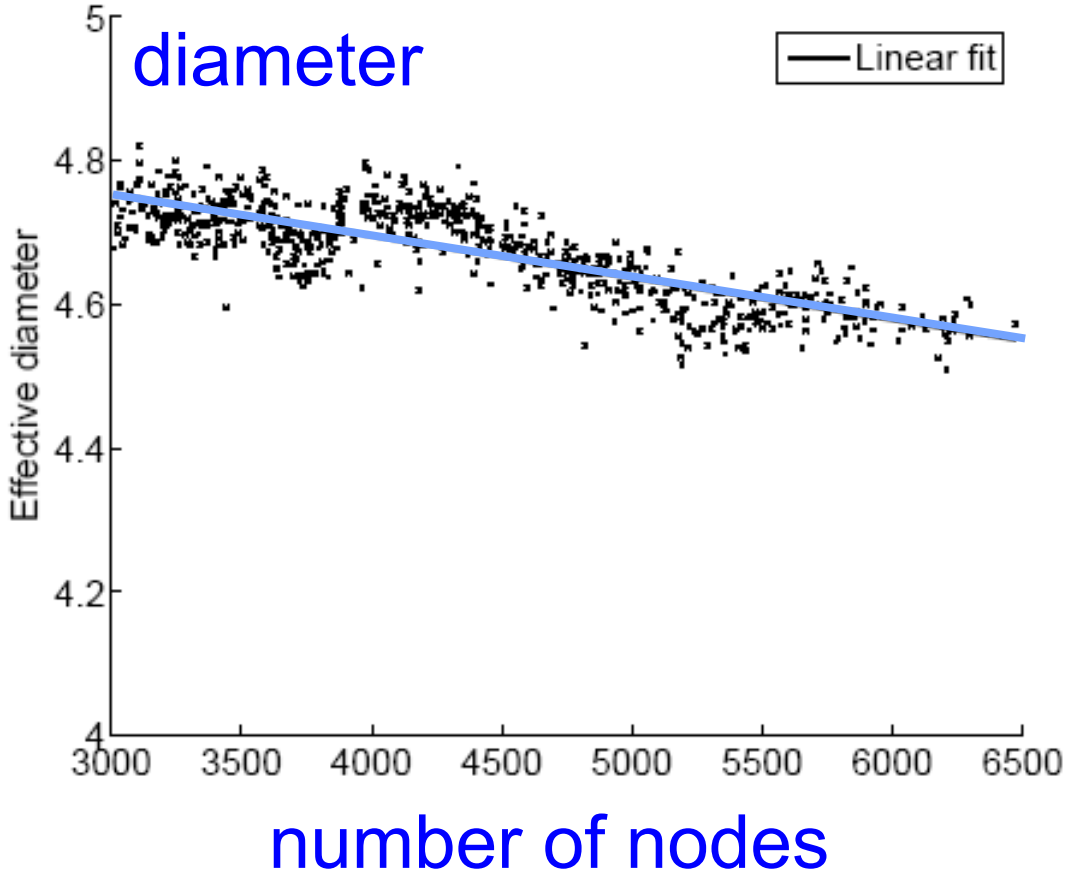
Diameter – ArXiv citation graph

- Citations among physics papers
- 1992 –2003
- One graph per year



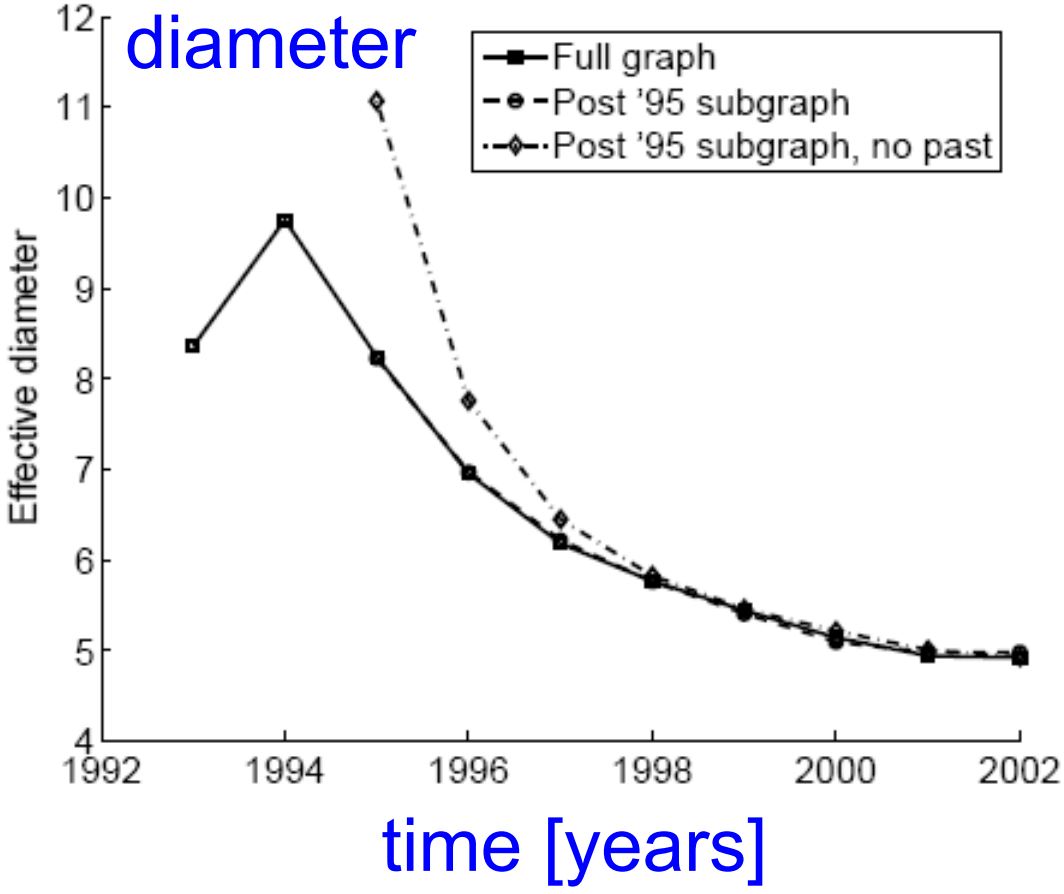
Diameter – “Autonomous Systems”

- Graph of Internet
- One graph per day
- 1997 – 2000



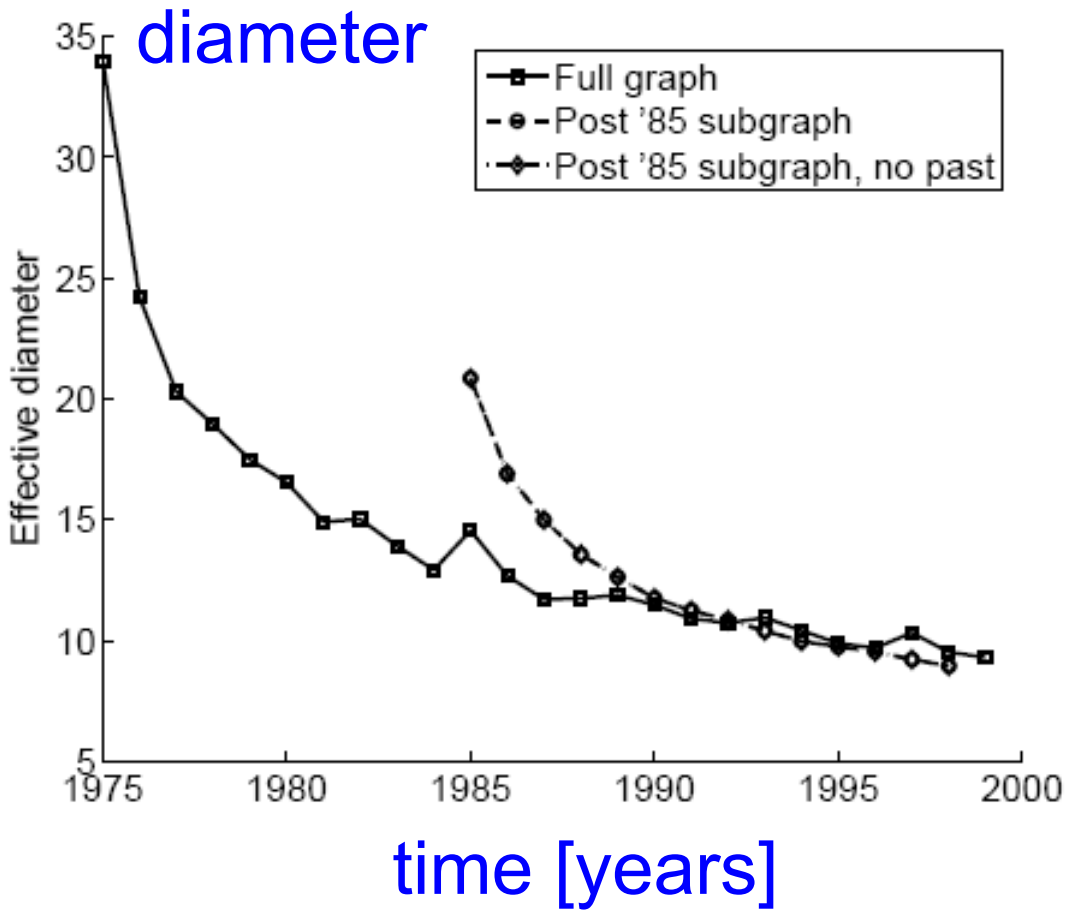
Diameter – "Affiliation Network"

- Graph of collaborations in physics – authors linked to papers
- 10 years of data



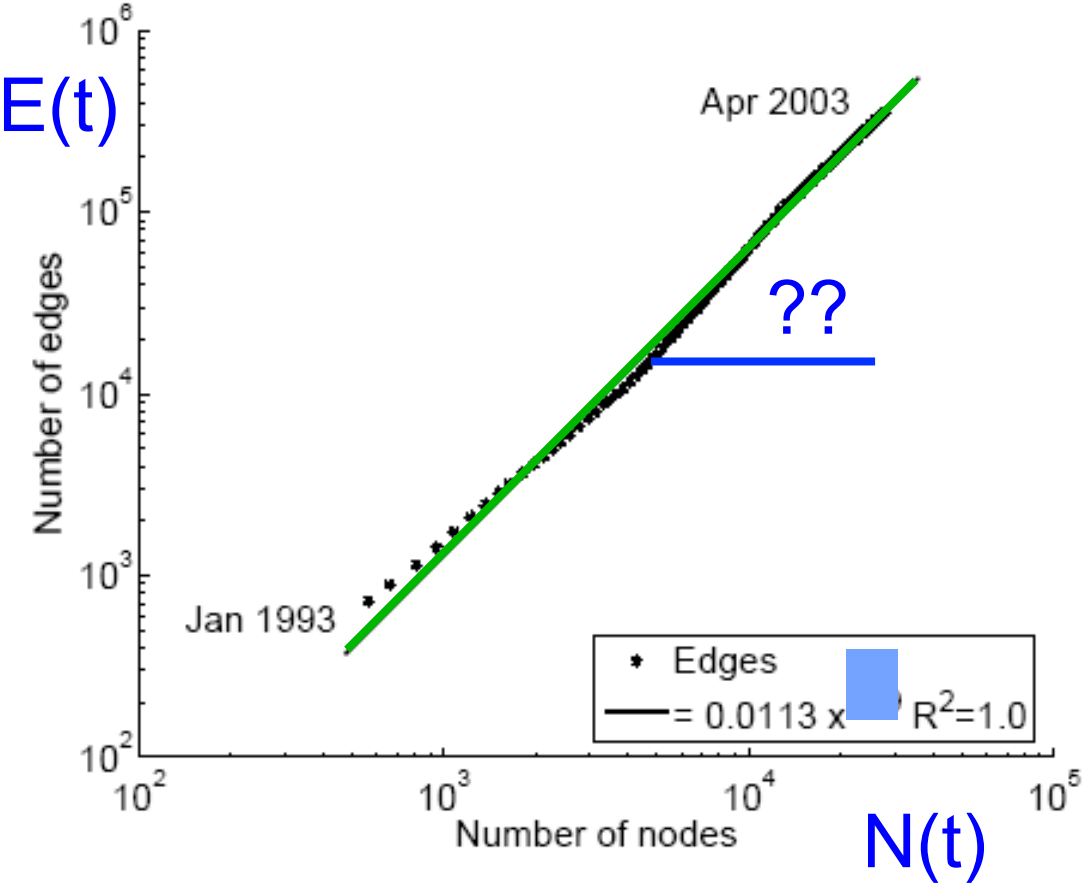
Diameter – “Patents”

- Patent citation network
- 25 years of data



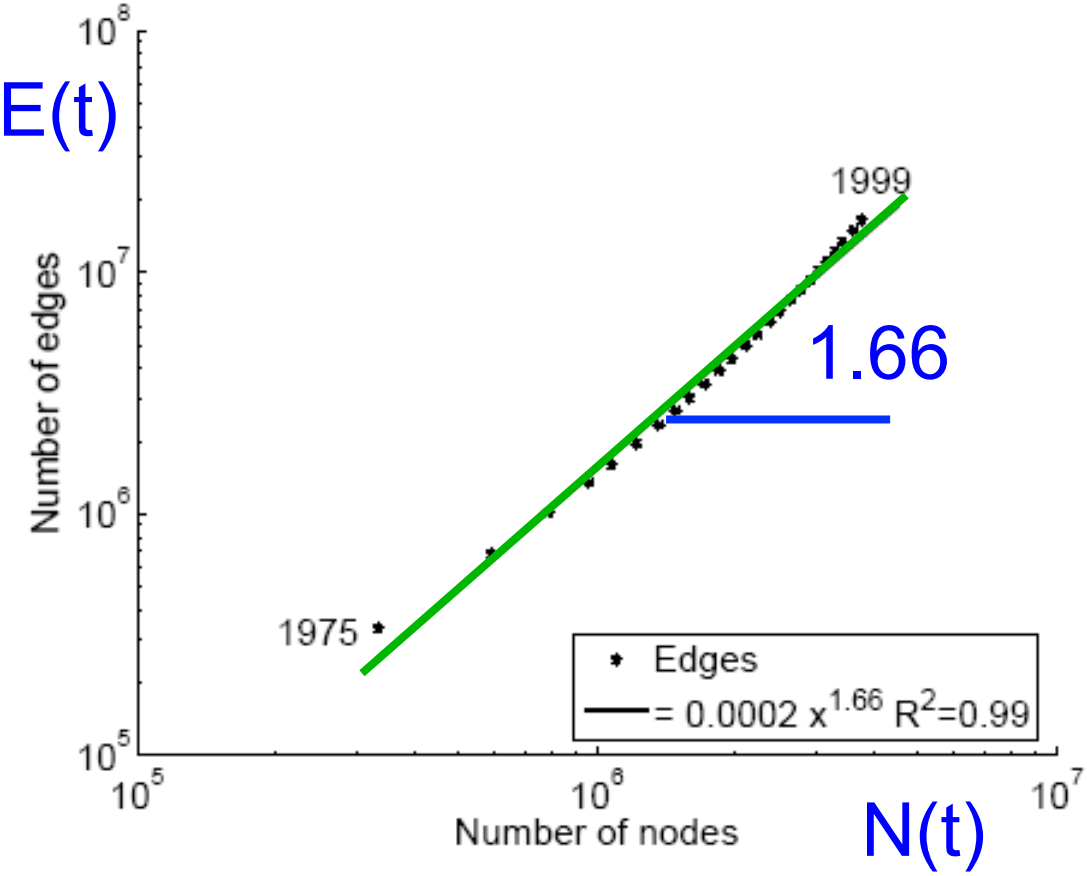
Densification – Physics Citations

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations



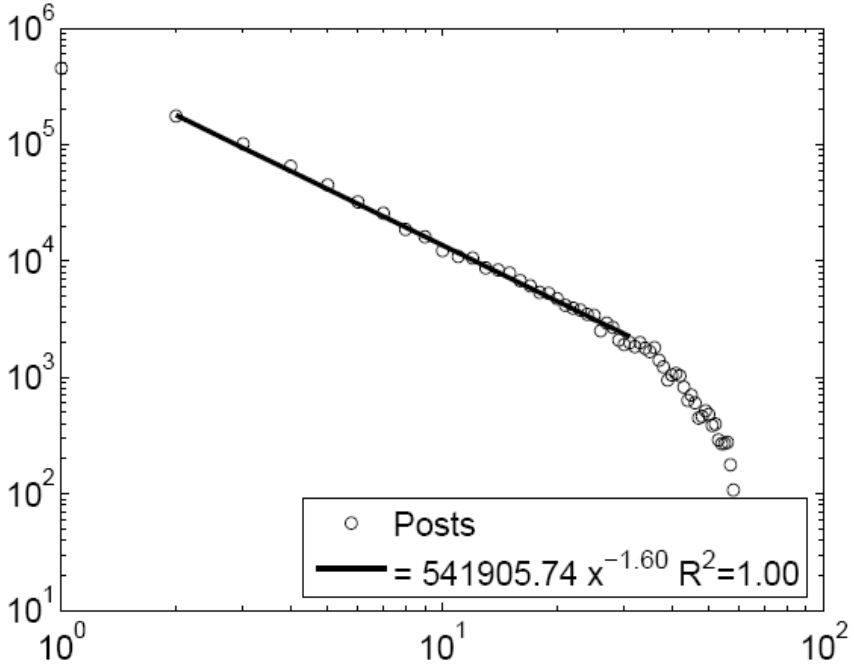
Densification – Patent Citations

- Citations among patents granted
- 1999
  - 2.9 million nodes
  - 16.5 million edges
- Each year is a datapoint



Popularity over time

in links (**log**)



days after post (**log**)

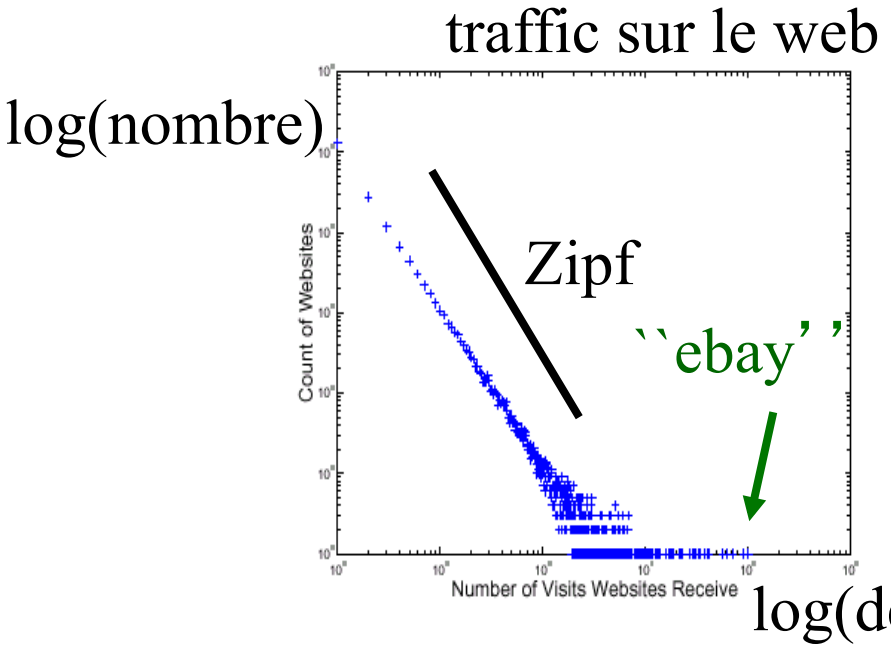
Post popularity drops-off ?

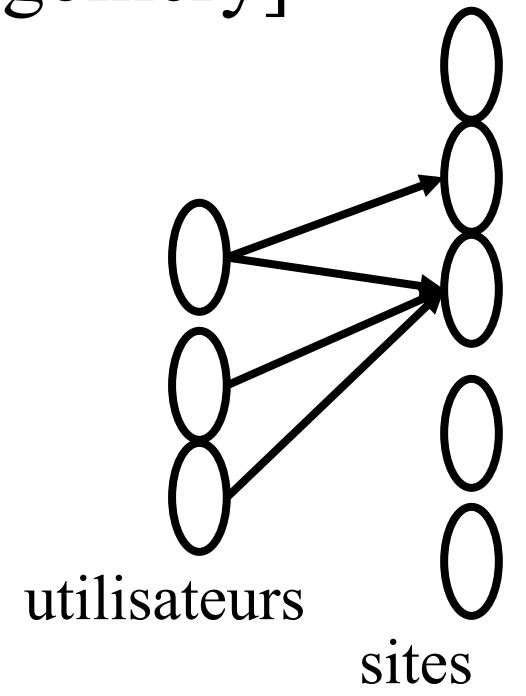
POWER LAW!

Exponent?

Nombre vs. popularité

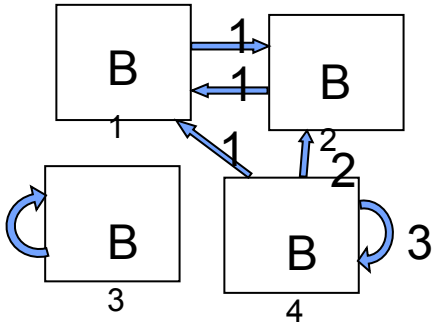
- "web hit counts" [w/ A. Montgomery]

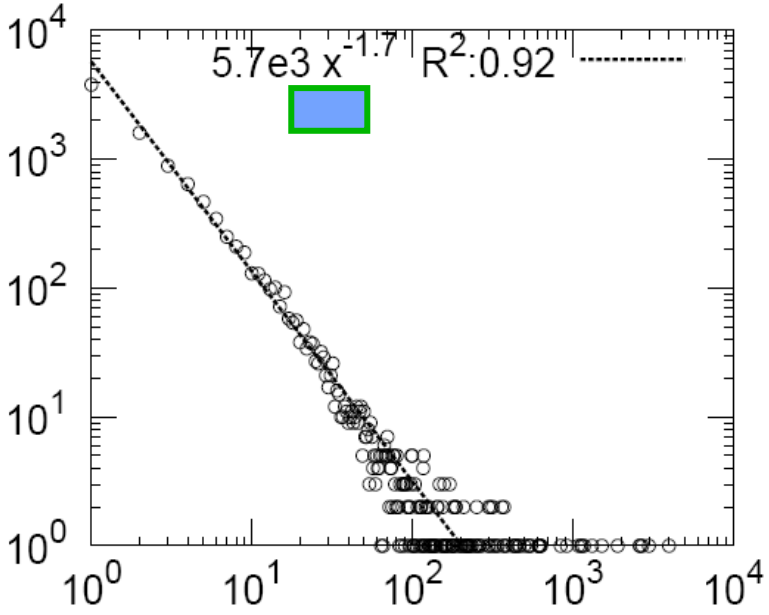


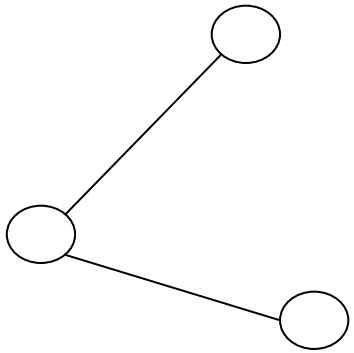


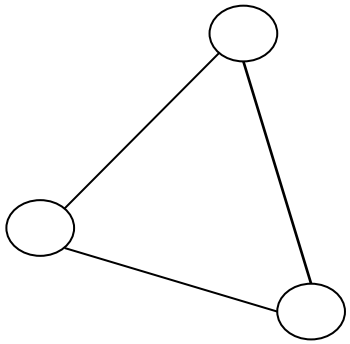
Degree distribution

44,356 nodes, 122,153 edges. Half of blogs belong to largest connected component.









Recherche de triangles

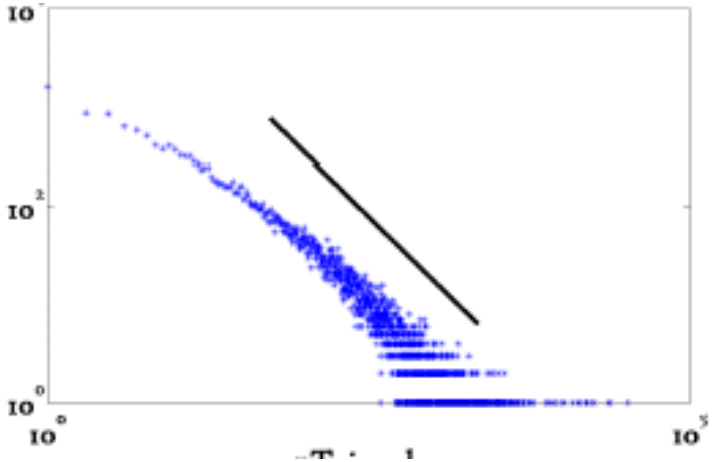
Les vrais réseaux sociaux devraient comporter beaucoup de triangles:
« les amis de mes amis sont mes amis »

DÉTECTION DE MOTIFS CARACTÉRISTIQUES

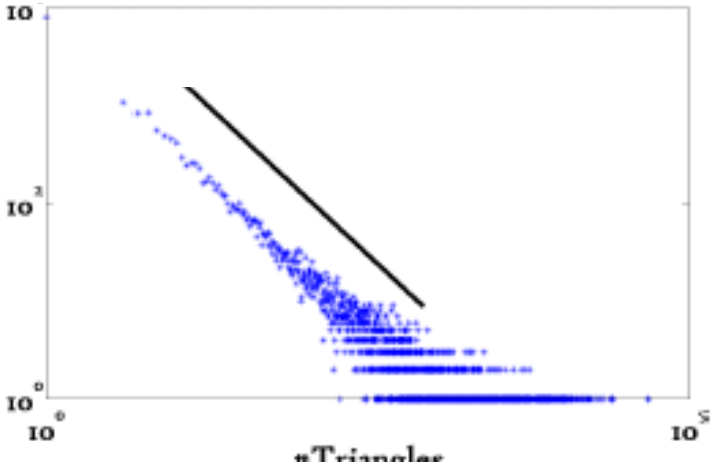
Loi de distribution des Triangles – 1

[Tsourakakis ICDM 2008]

HEP-TH



Epinions



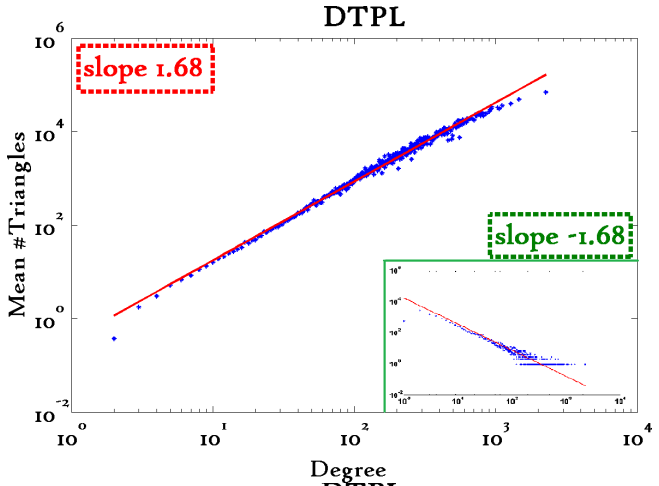
Abscisse: nbre de triangles auxquels un nœud appartient

Ordonnée: nombre de tels nœuds

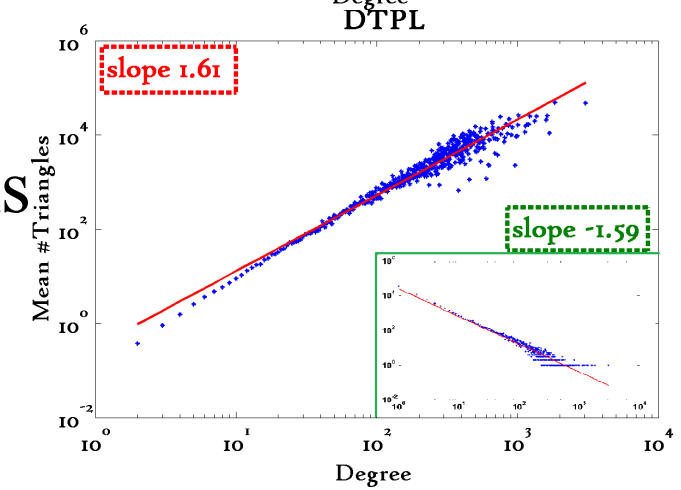
Loi de distribution des Triangles – 2

[Tsourakakis ICDM 2008]

Reuters



Epinions



Abscisse: degré des nœuds

Ordonnée: nombre moyen de triangles

Extraction de motifs fréquents: nombre d'occurrences (support) supérieur à un nombre minimal

EXTRACTION DE MOTIFS

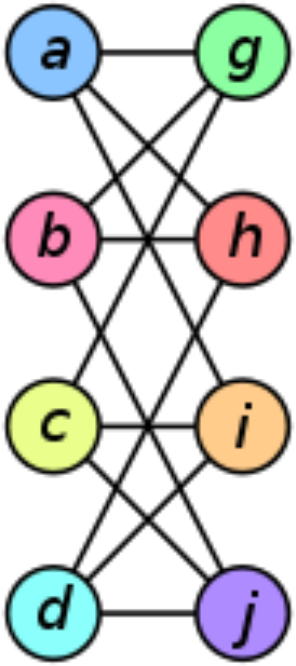
Terminologie 1 – rappels

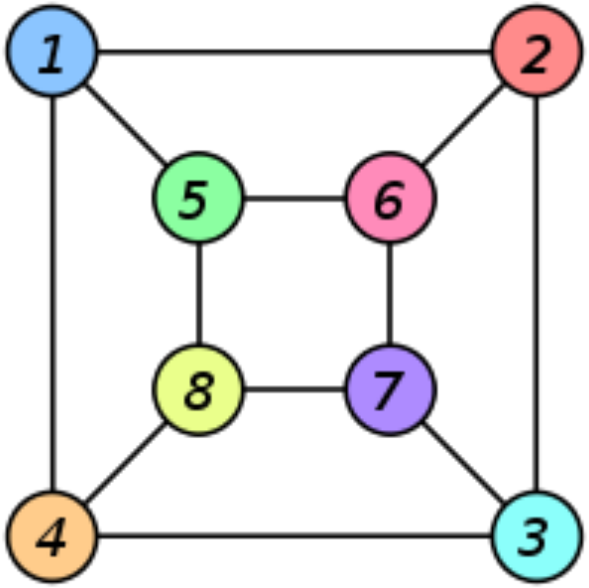
- Un graphe G(V,E) est composé de deux ensembles
  - V: ensemble d'**arc**/d'**arêtes**
  - E: ensemble de **nœuds/sommets**
- On considère ici des graphes non orientés et étiquetés
  - L_v: ensemble d'étiquettes d'arêtes
  - L_E: ensemble d'étiquettes de sommets
- Les étiquettes n'ont pas besoin d'être uniques
  - par exemple, nom des éléments dans une molécule

Terminologie 2 – rappels

- Un graphe est dit ***connecté*** s'il existe un chemin entre toute paire de nœuds
- Un graphe G_s (V_s, E_s) est un ***sous graphe*** d'un autre graphe G(V, E) si et seulement si
  - V_s est un sous-ensemble de V et E_s un sous-ensemble de E
- Deux graphes $G_1(V_1, E_1)$ et $G_2(V_2, E_2)$ sont ***isomorphes*** si leur topologie est identique
  - Il existe un appariement de V_1 vers V_2 tel que chaque arête de E_1 est appariée à une arête unique de E_2 et vice-versa
- Recherche d'isomorphisme entre deux graphes $G_1(V_1, E_1)$ et $G_2(V_2, E_2)$
  - Problème NP complet

Exemple d'isomorphisme de graphes





$f(a) = 1$

$f(b) = 6$

$f(c) = 8$

$f(d) = 3$

$f(g) = 5$

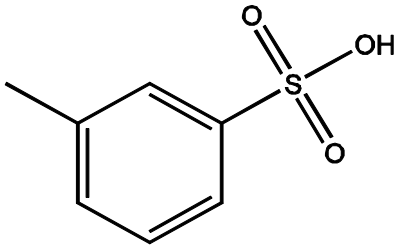
$f(h) = 2$

$f(i) = 4$

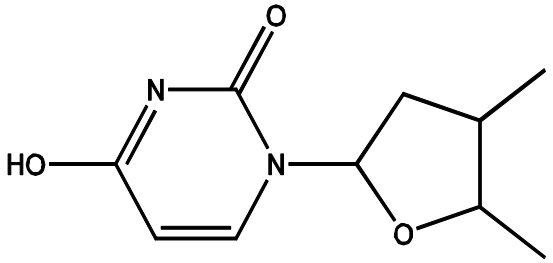
$f(j) = 7$

Exemple 1

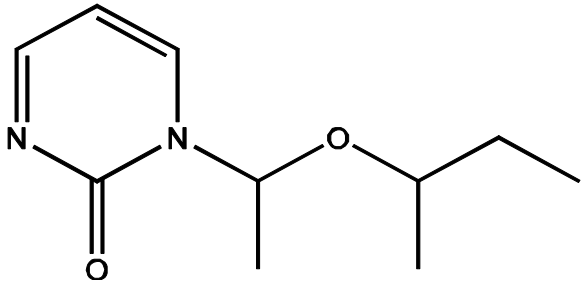
Ensemble de graphes



(T1)



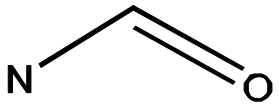
(T2)



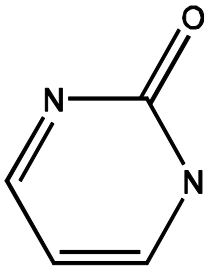
(T3)

Motifs fréquents
(MIN SUPPORT = 2)

(1)

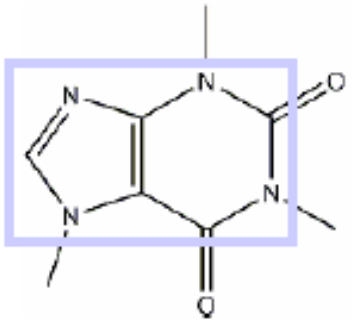


(2)

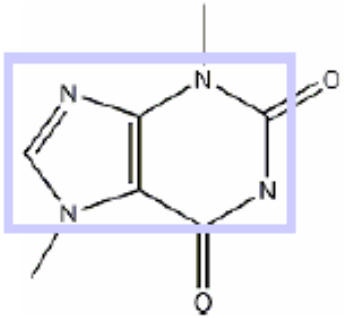


Exemple 2

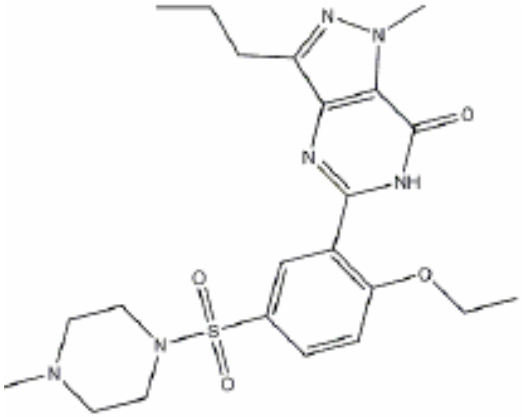
Ensemble de graphes



(a) caffeine

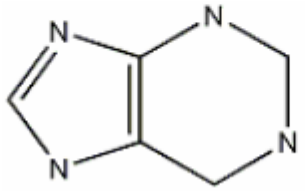


(b) diurobromine



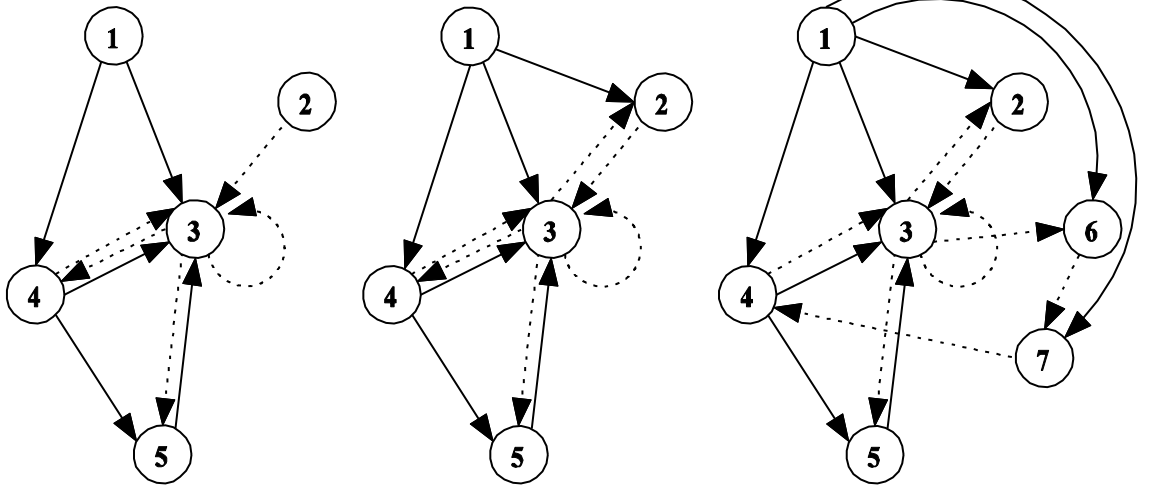
(c) viagra

Motifs fréquents
(MIN SUPPORT = 2)

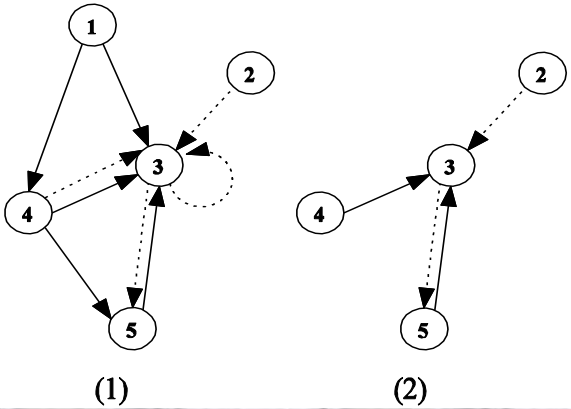


Exemple 3

Ensemble de graphes



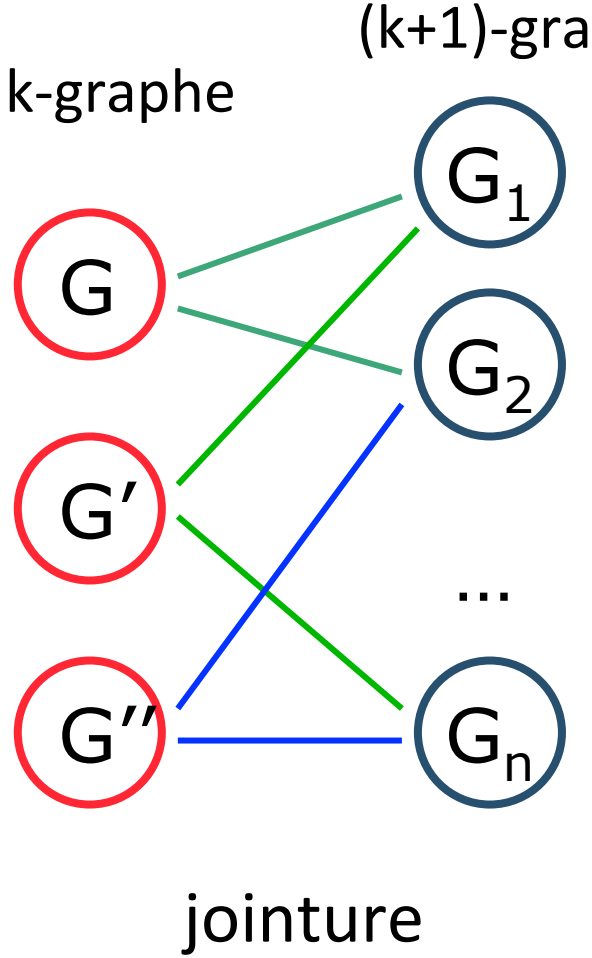
Motifs fréquents
(MIN SUPPORT = 2)



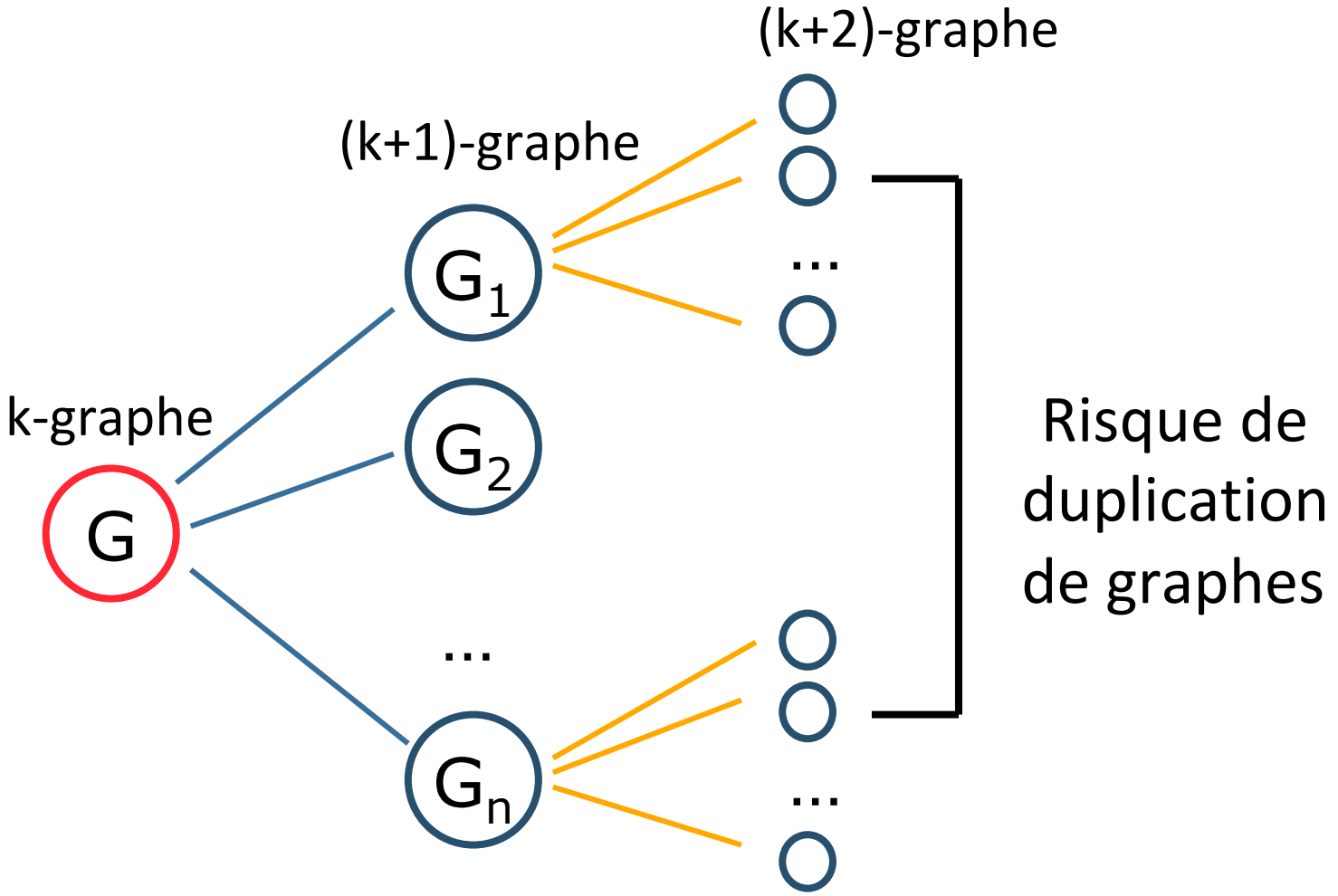
Méthodes d'extraction de motifs

- Approches « Apriori » (jointure)
- Approche basée sur la croissance de motifs

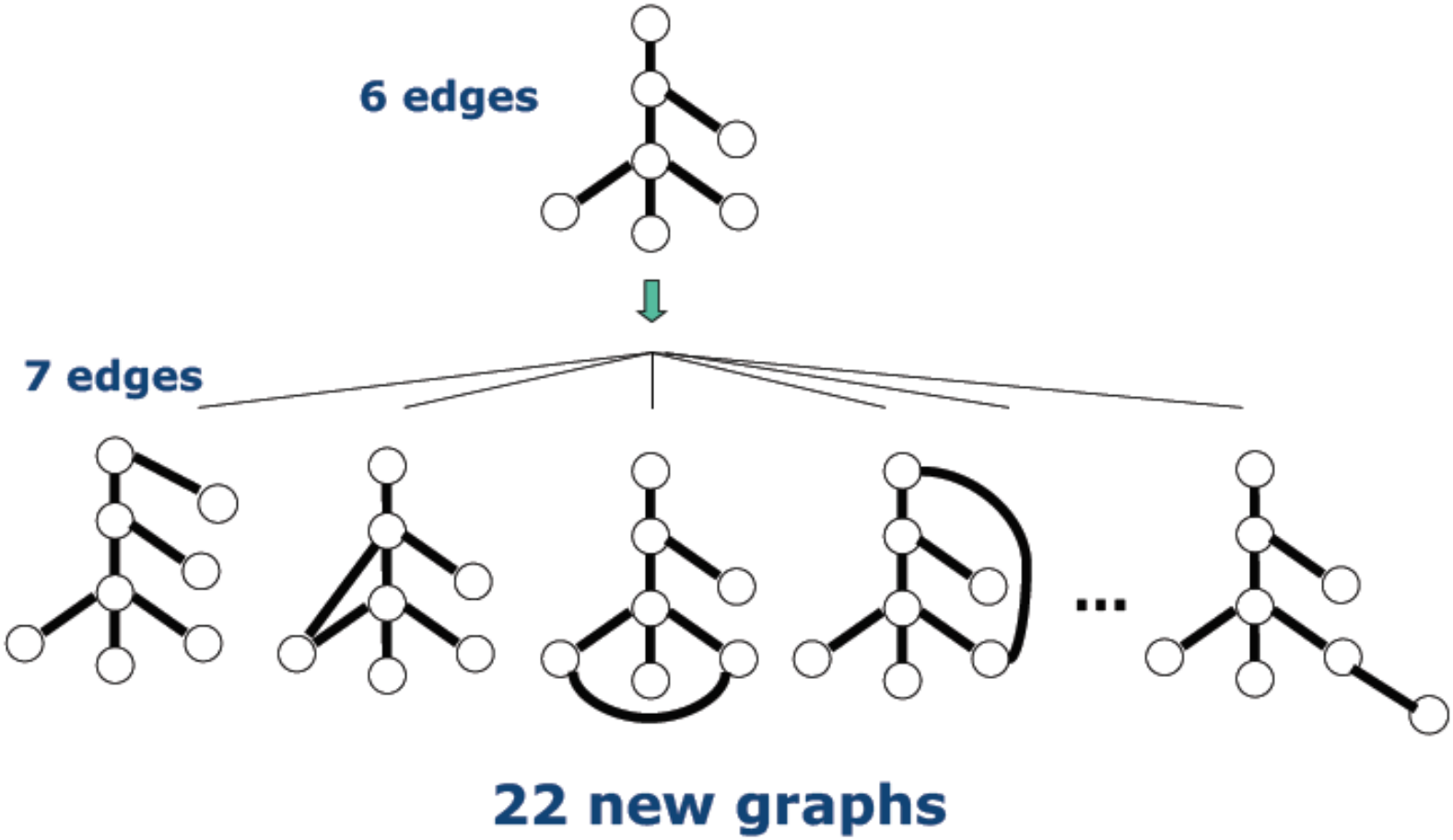
Approche Apriori



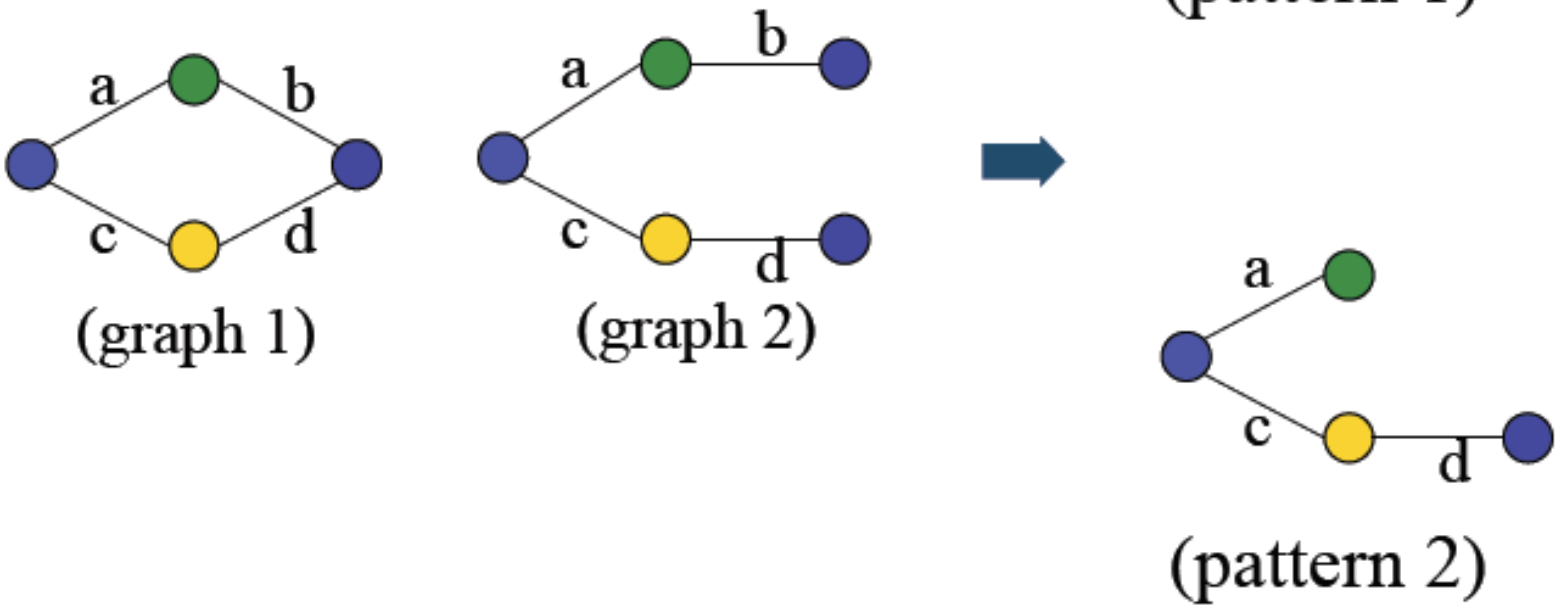
Approche par croissance de motifs



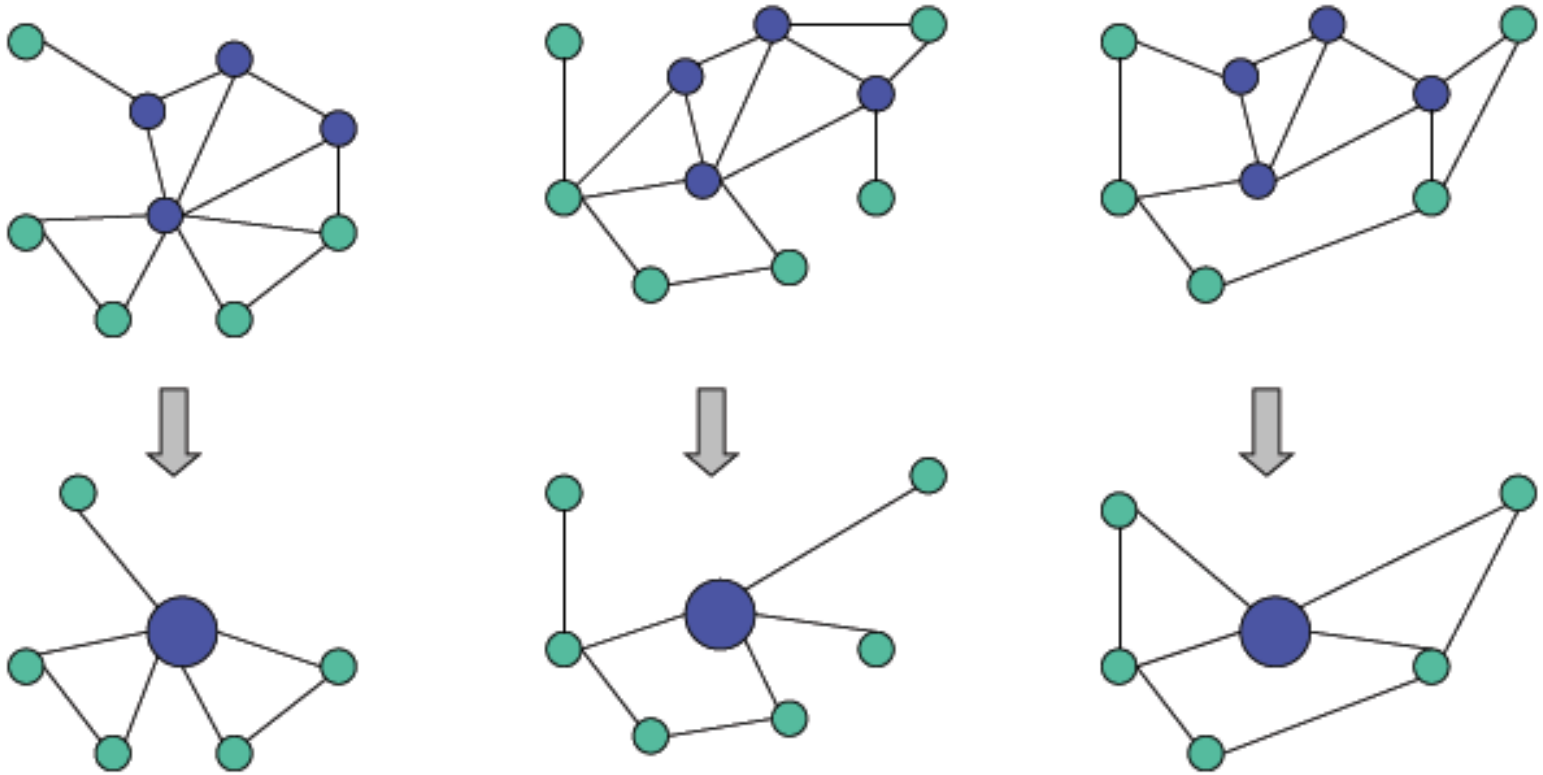
Libre génération



Cas délicats



Compression de graphes



Recherche de sous-graphes communs

- **Entrée:** (*D, minSup*)
  - Ensemble de graphes étiquetés $D=\{T_1, T_2, \ldots, T_N\}$
  - Support minimum *minSup*
- **Sortie:** (tous les sous-graphes communs).
  - Un sous-graphe est dit fréquent si c'est un sous-graphe d'au moins $minSup \cdot |D|$ éléments de *D*.
  - Chaque sous-graphe est connecté.