Mesure et Analyse de l’activité (pédophile) sur les réseaux P2P

Ludovic Denoyer - Equipe MALIRE
Guillaume Valadon – Equipe Complex Networks

UPMC PARIS UNIVERSITAS

cnrs

Contexte

- Etude des réseaux d'échange / réseaux sociaux
  - Réseaux large scale

- Cas d'application: Projet MAPAP – Analyse de la pédophilie sur les réseaux P2P

- Trois axes de recherche:
  - Mesure sur les réseaux
    - Influence de la mesure
  - Analyse de la topologie des réseaux
    - Analyse de communautés
  - Analyse de contenu
    - Analyse conjointe structure/contenu

Equipe Complex networks

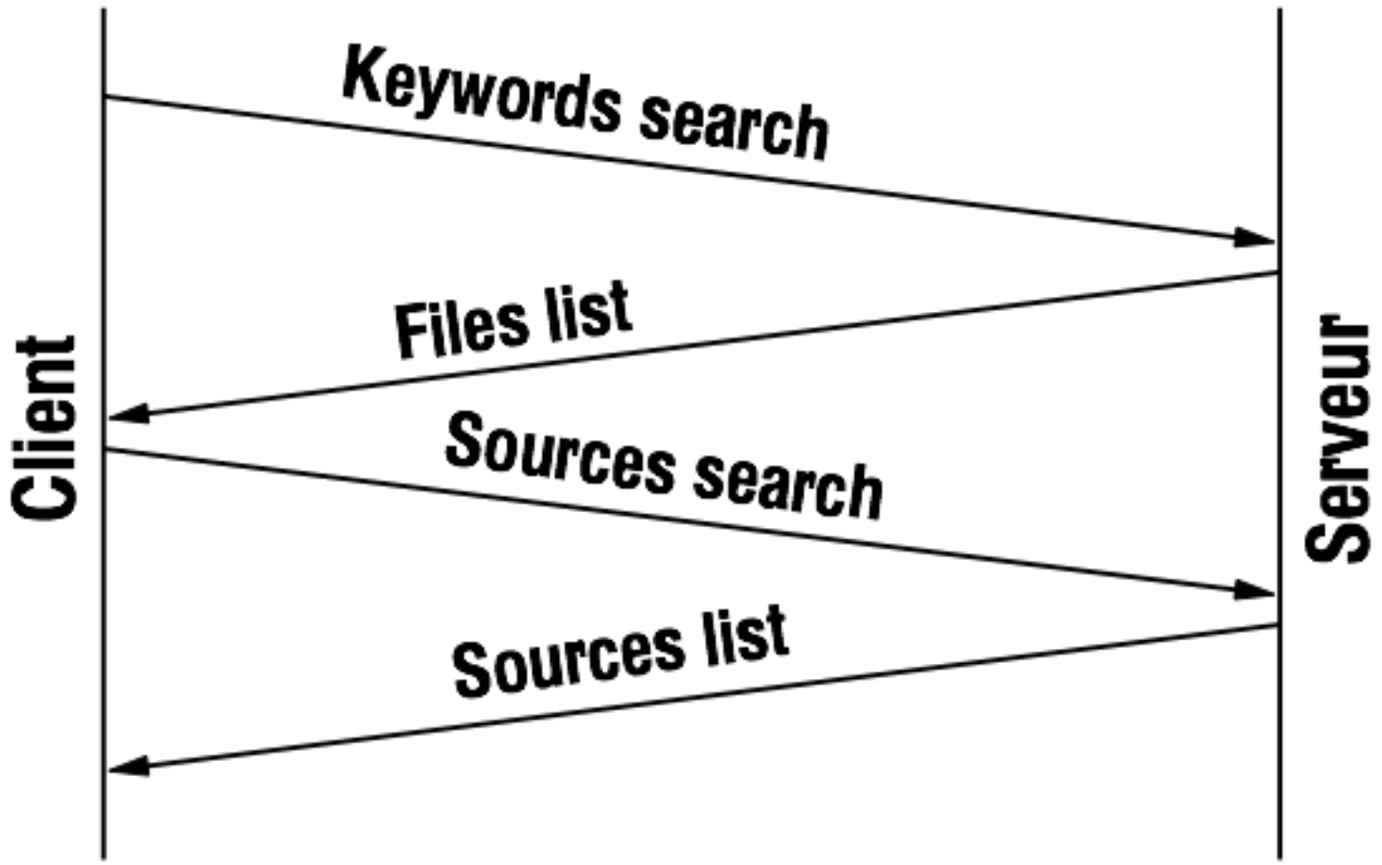
Equipe MALIRE

Problématiques de mesure

- Etude des échanges sur les réseaux P2P
  - Diffusion de fichiers
  - Communautés d'intérets
  - Popularité
- Motivations
  - Développer de nouveaux protocoles
  - Détection du contenu caché
  - Simulation de protocole et d'échanges
  - Analyser un certain type d'activité (pédophile)

Exemple: Echanges eDonkey

1. Entre clients: telechargement de fichiers
2. Entre serveurs: données statistiques
3. Client-serveurs: Recherche de fichiers

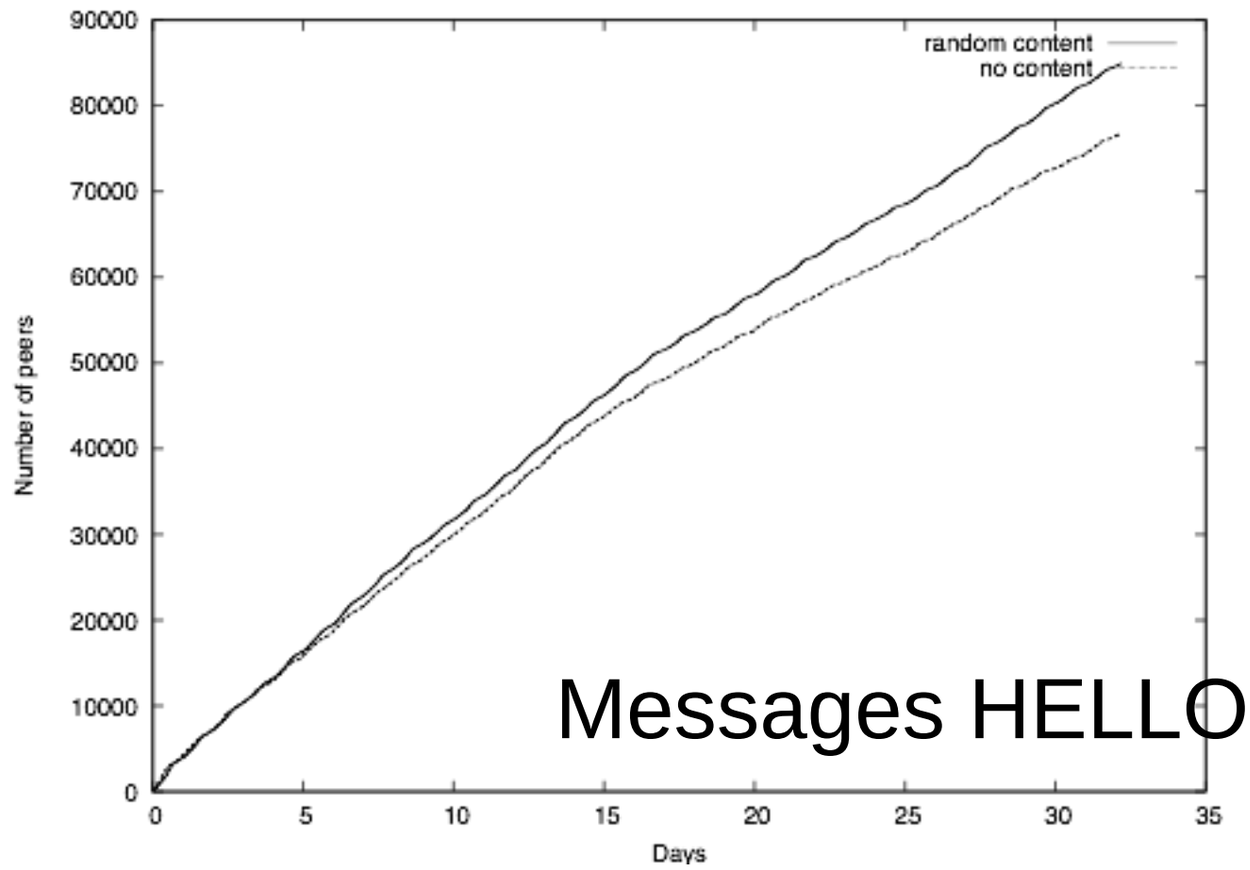


Méthodologie de collecte

- Différents types:
  - Collecte passive
  - Collecte active

- Problèmes liés aux mesures:
  - Efficacité de la collecte
  - Qualité des observations
  - Paramètres utilisés pour les mesures

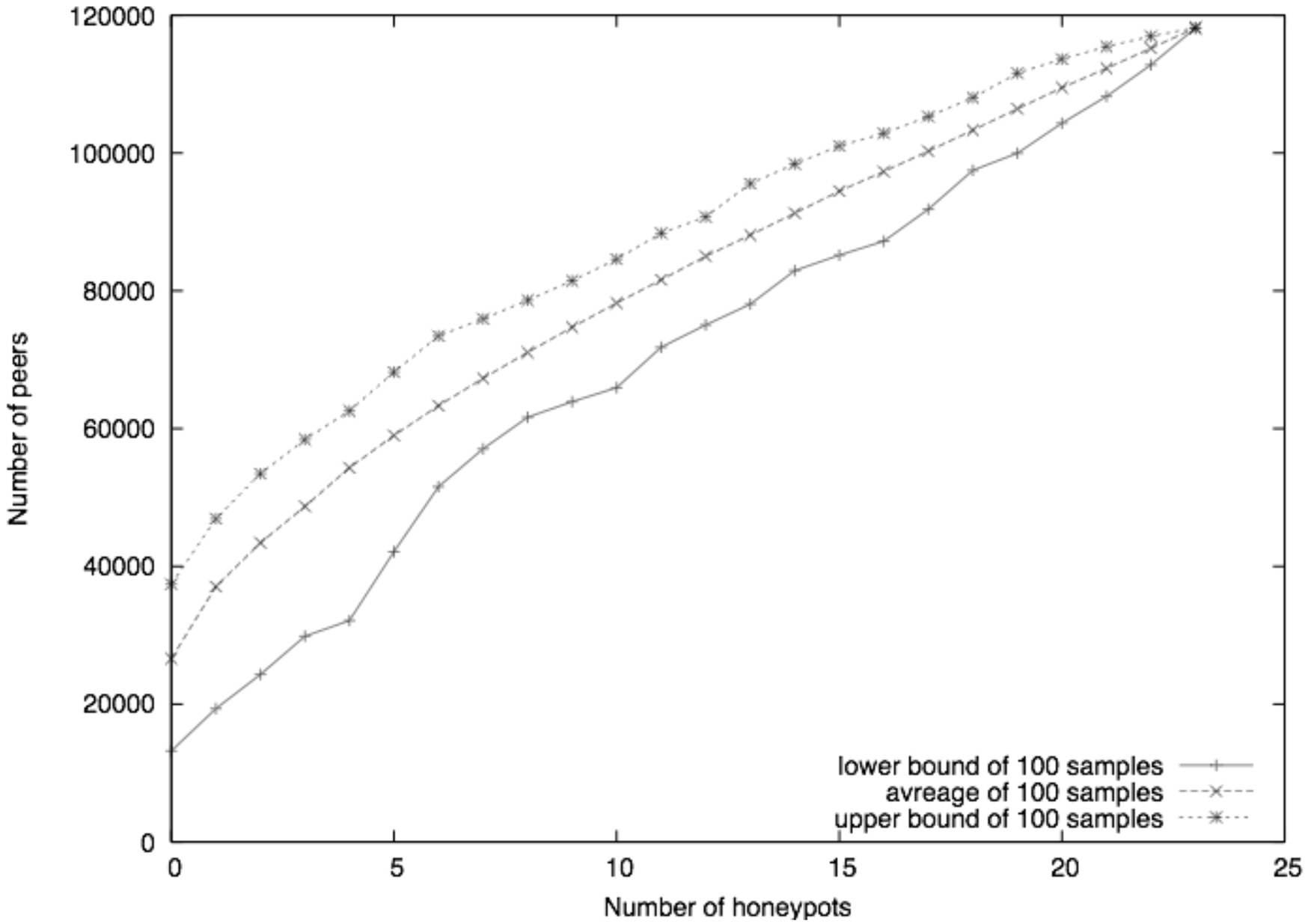
Paramètres : contenu vide ou aléatoire



- Problème légal : on ne peut pas envoyer de contenu
- But : voir un maximum de clients

- On voit plus de clients avec la stratégie aléatoire

Paramètres : nombre d’agents



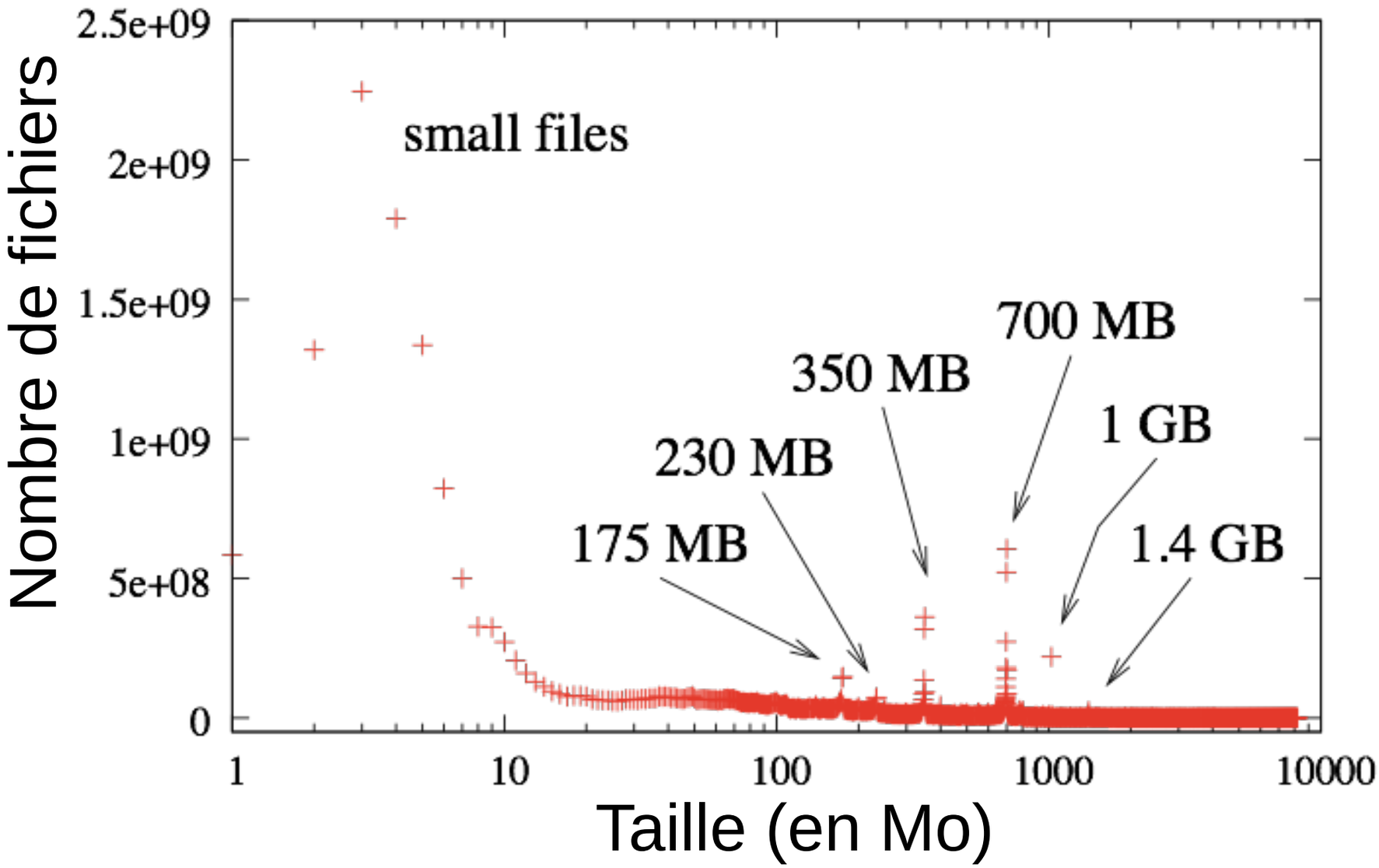
- On voit plus de clients lorsque le nombre d’agen augmente

Capture passive sur eDonkey

- 10 semaines de mesures sur un serveur
- ~500 GB de XML compressé
- ~ 10 milliards de messages
- ~ 90 millions de clients
- ~ 280 millions de fichiers différents

- Données anonymisées disponibles ici : http://antipaedo.lip6.fr

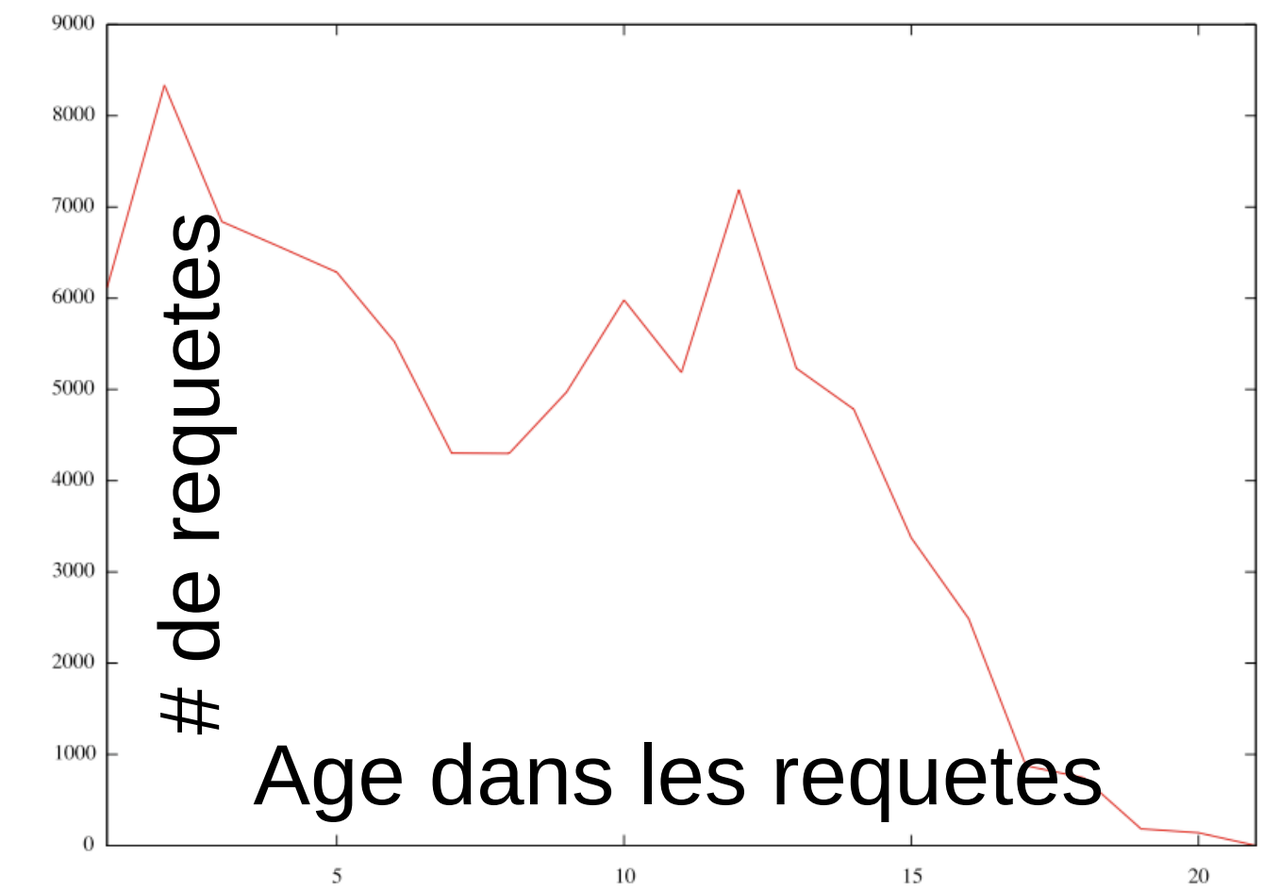
Analyse des traces collectées

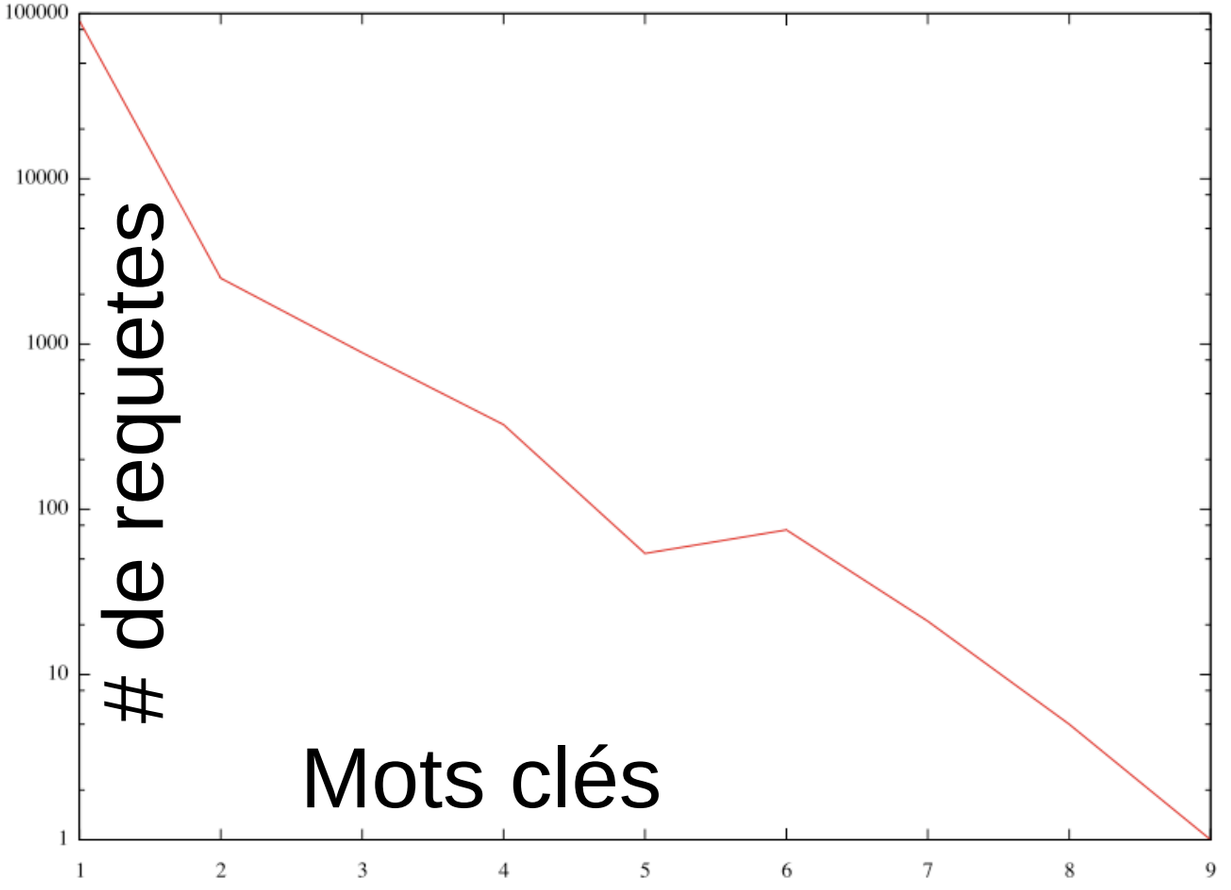


- Taille d'un CD-ROM et fractions (1/2, 1/3, et 1/4)
  - relation avec les tailles des supports classiques

Contenu pédophile

	Requêtes	Nom de fichiers
Mots clés pédophiles	94288	7293
Ages (XYyo, XYyr)	56672	10026





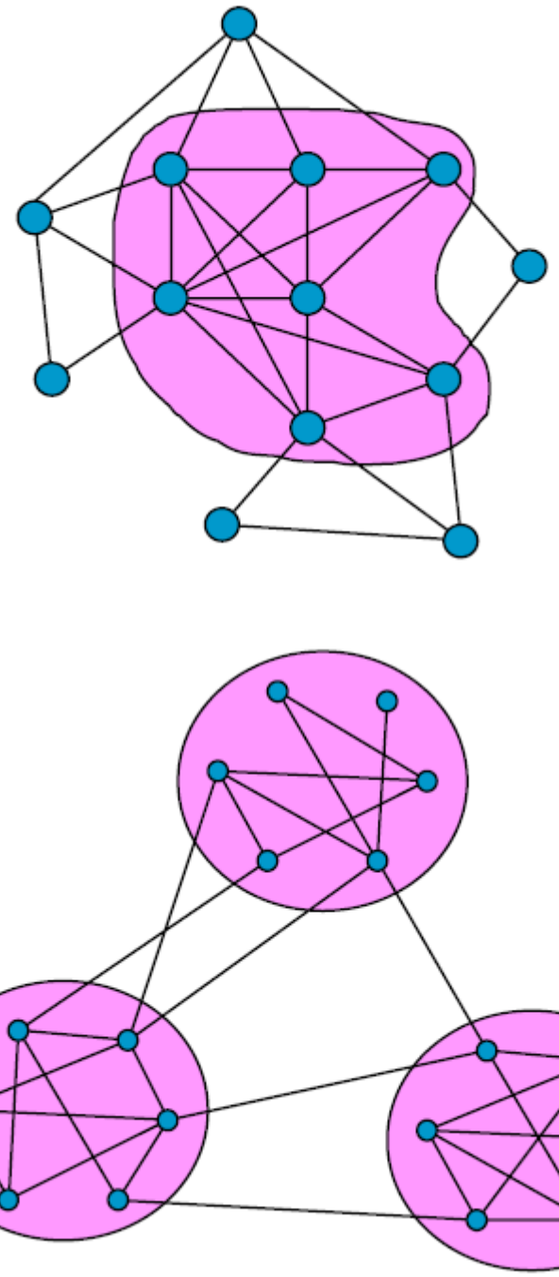
Analyse de topologie et de contenu

Analyse de topologie et de contenu

- Nécessité d'analyser les données collectées de manières « rusée »
  - Non supervisé: Détection automatique de communautés (clusters d'utilisateurs, de fichiers, ….)
  - Supervisé : Etiquetage des données
    - Détection d'utilisateurs/mots-clefs pédophiles

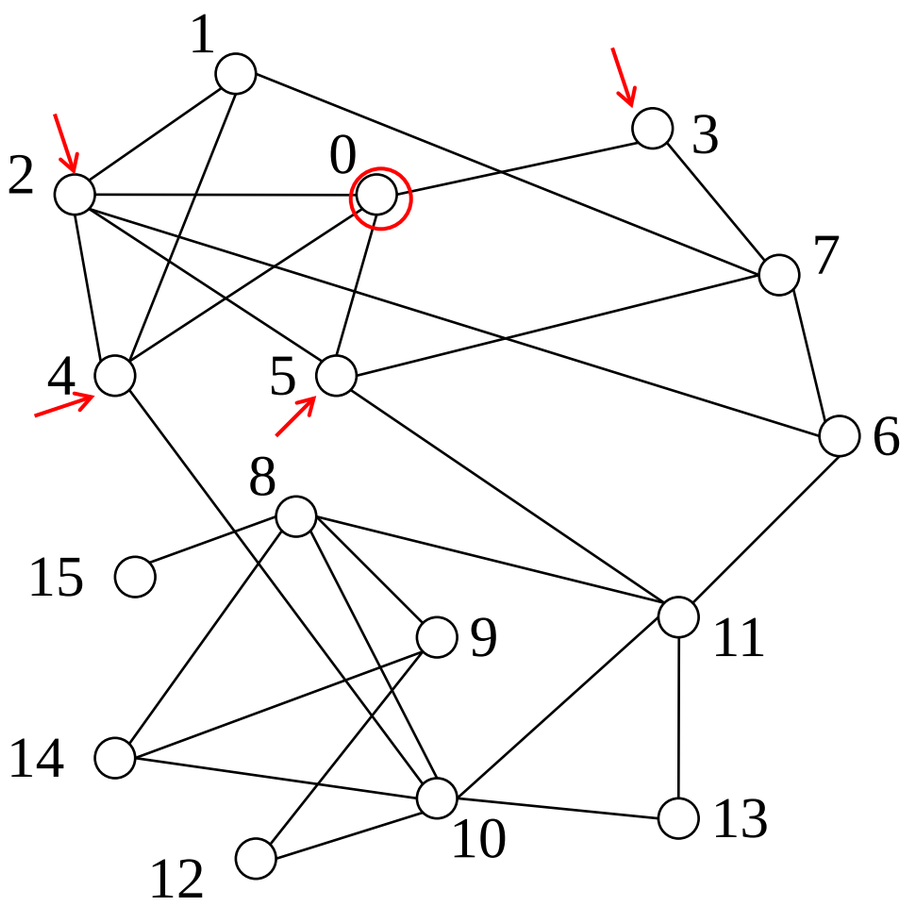
- Utilisation simultanée de l'information de structure et de contenu

Analyse de communautés topologiques



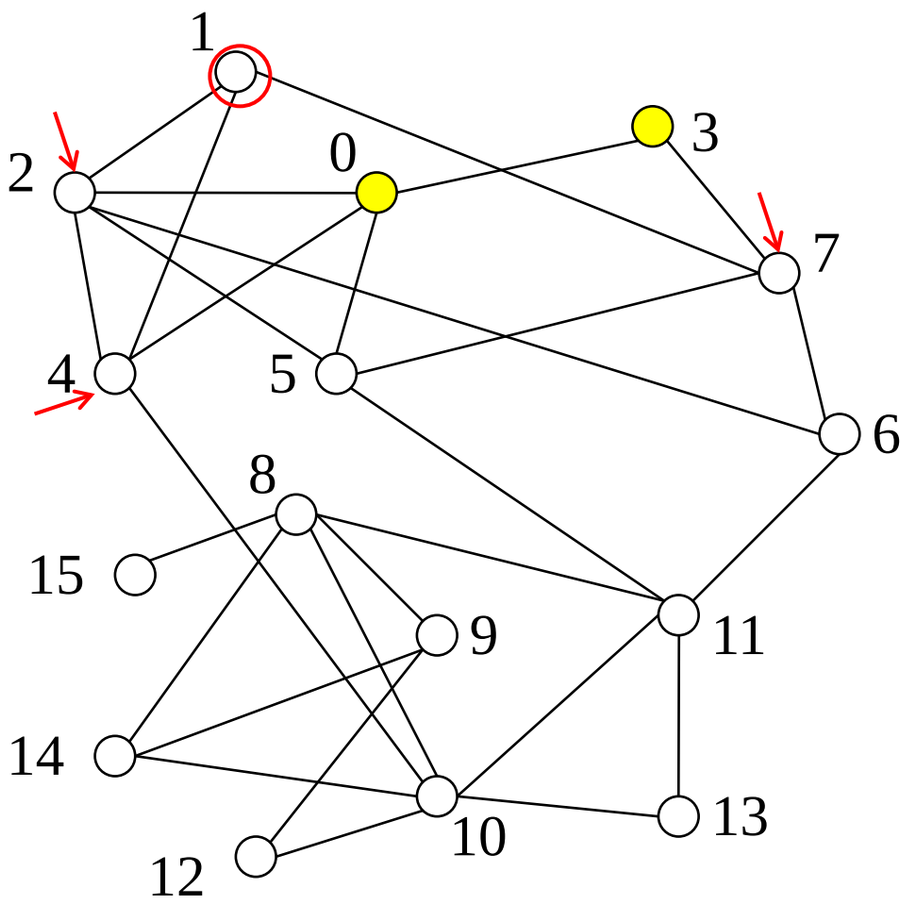
- Détection de composantes:
  - Dont les éléments sont fortement connectés
  - Dont les éléments sont faiblement connectés avec ceux des autres communautés
- Différentes approches
  - Clustering hiérarchique
  - Approche par division
  - Approche par méthode de monte-carlo
- Nécessité de développer des méthodes permettant le traitement de très grandes masses de données

Illustration

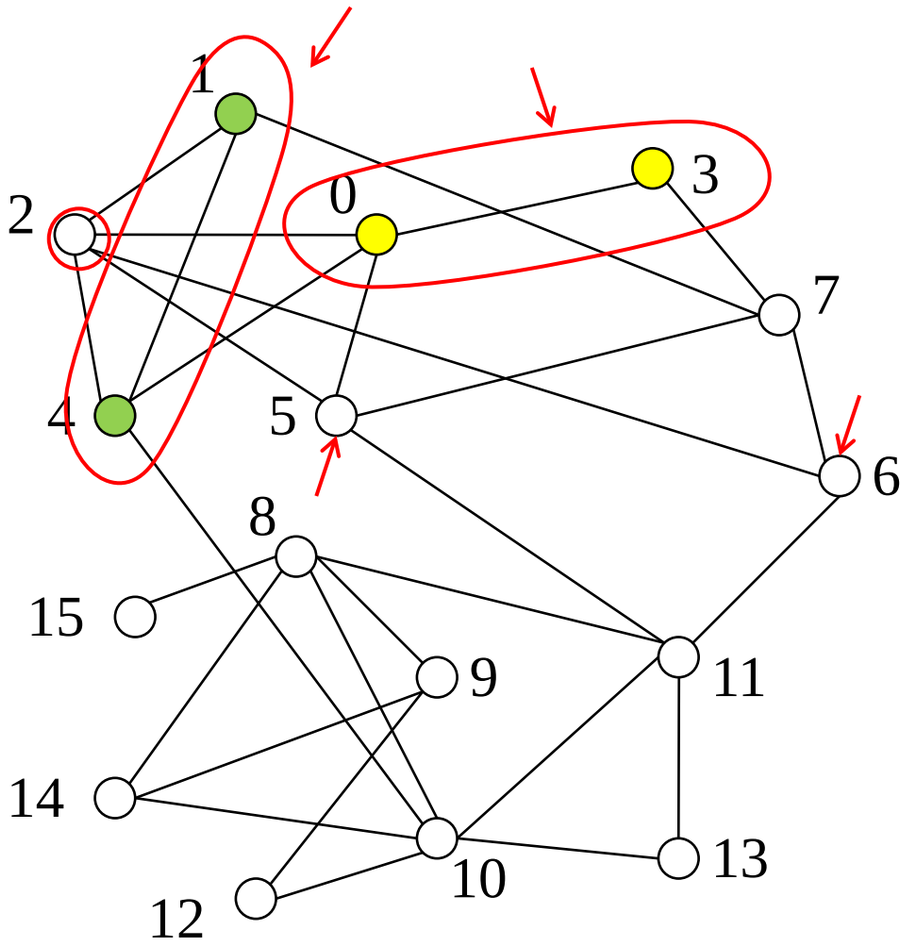


Pass 1 – Iteration 1
Each node belongs to an atomic community

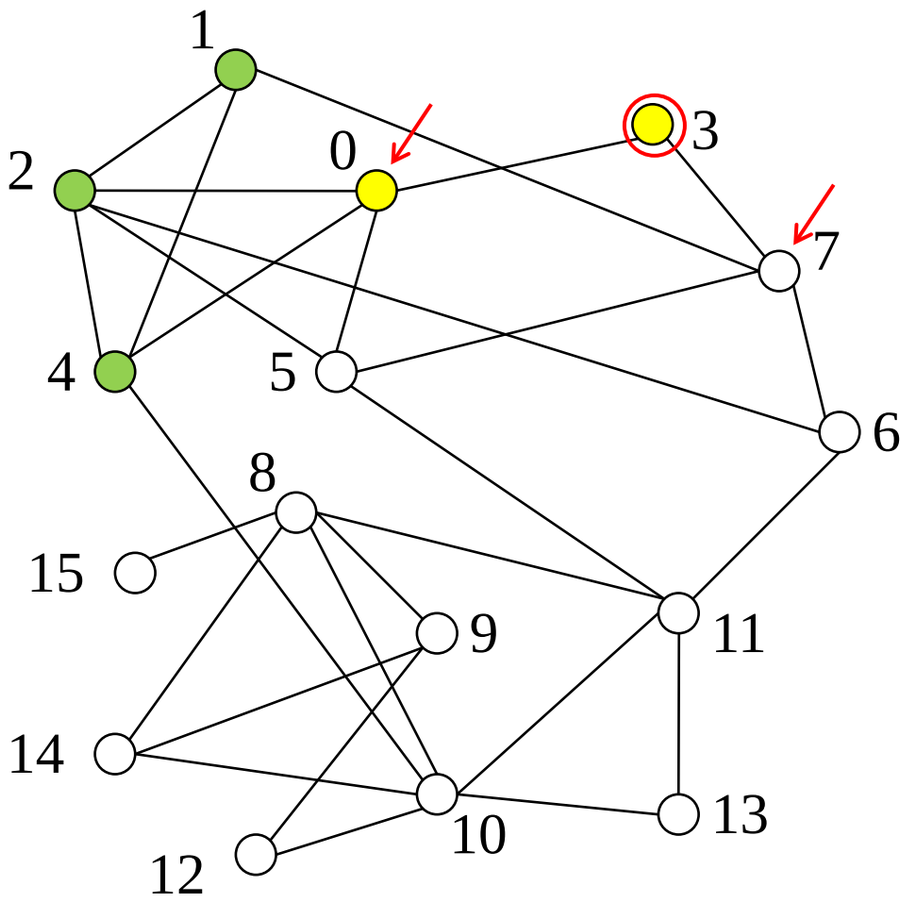
Illustration



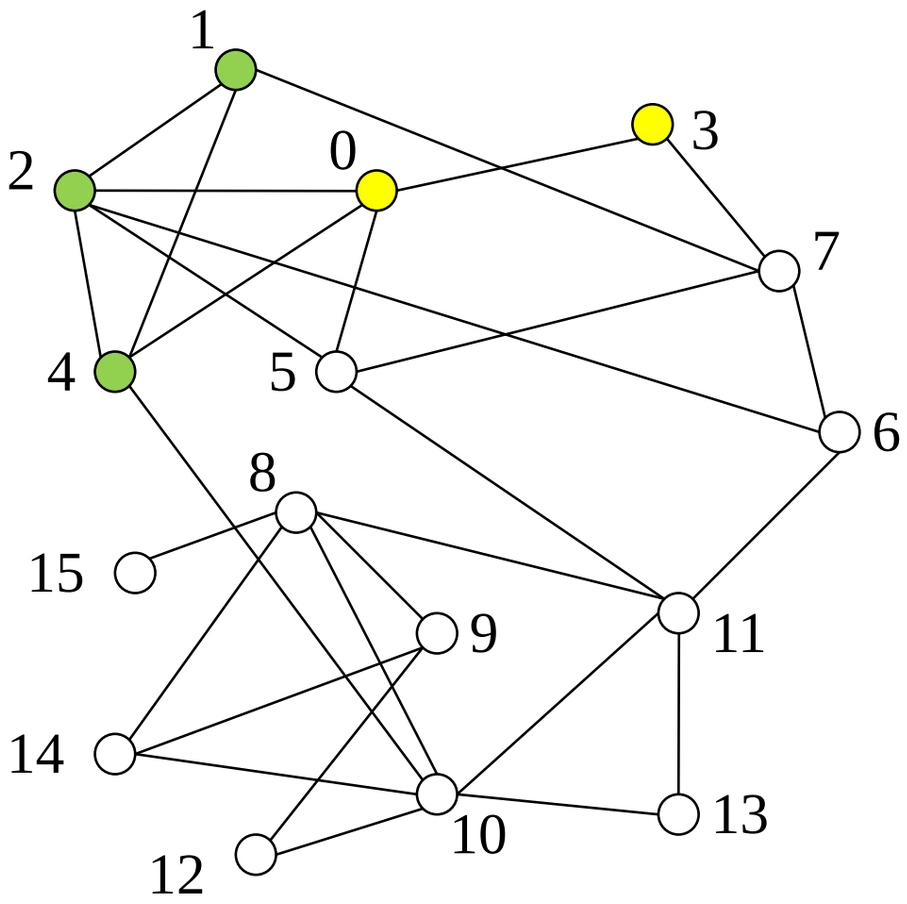
Illustration



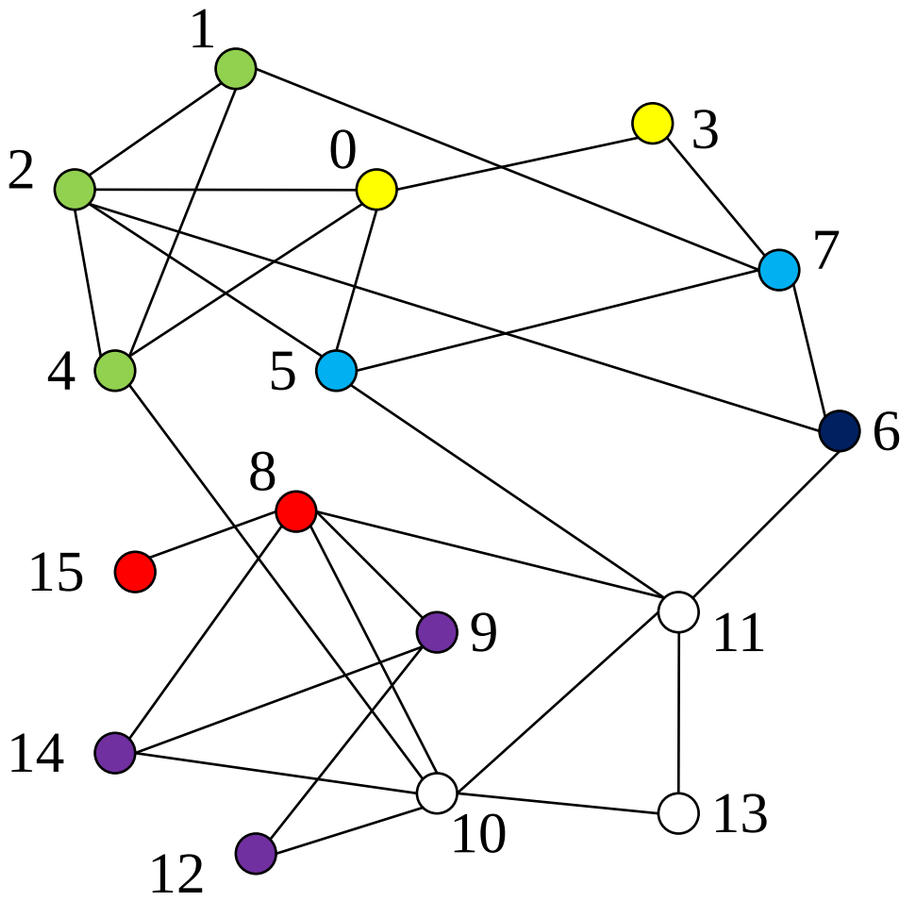
Illustration



Illustration



Illustration

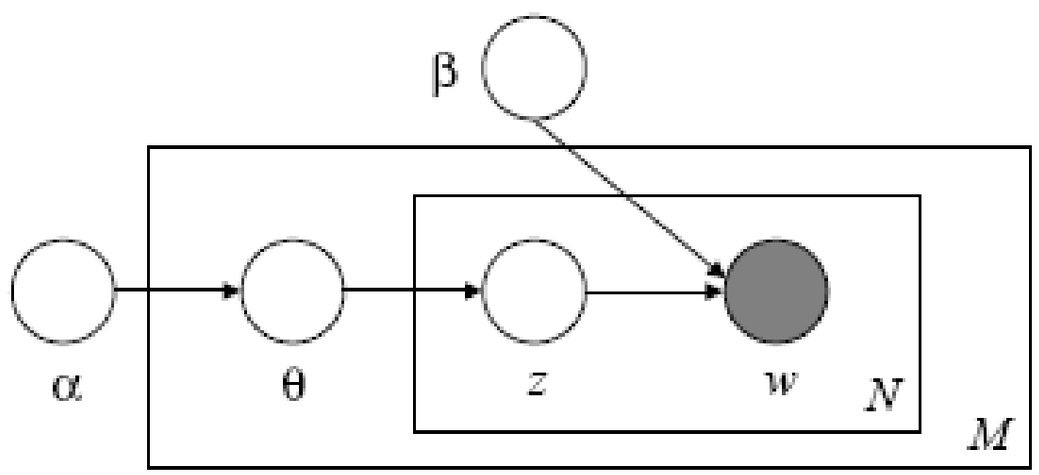


Analyse de communautés topologiques

- Limites: pas d'utilisation du contenu
  - Pas de détection de communautés thématiques
  - Deux communautés séparées peuvent avoir le même contenu
  - Pas de prise en compte de la description des éléments (utilisateurs par exemple)

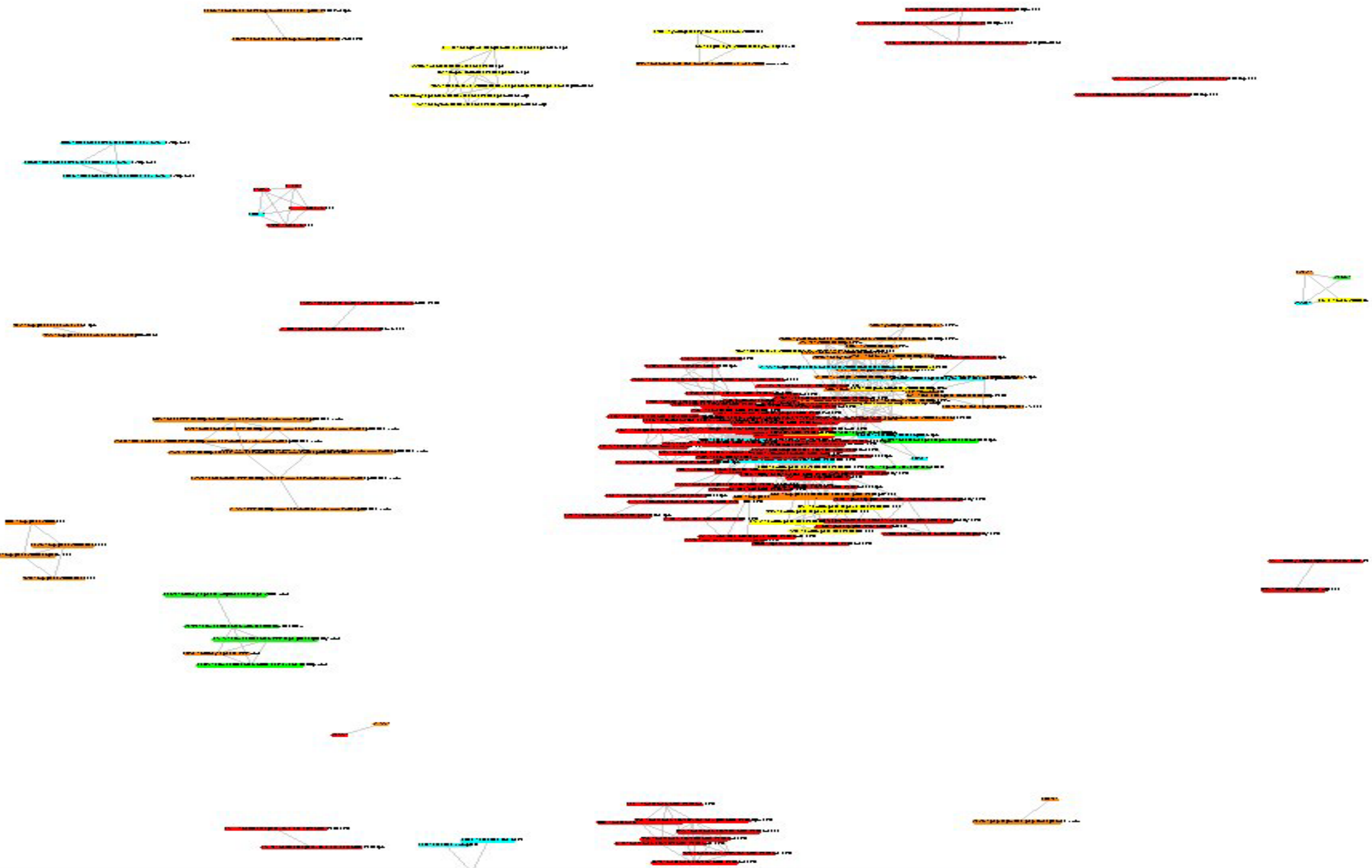
Détection de communautés Structure+Contenu

- Méthodes à base de modélisation par variable latentes
  - PLSA, LDA, méthodes spécifiques au P2P

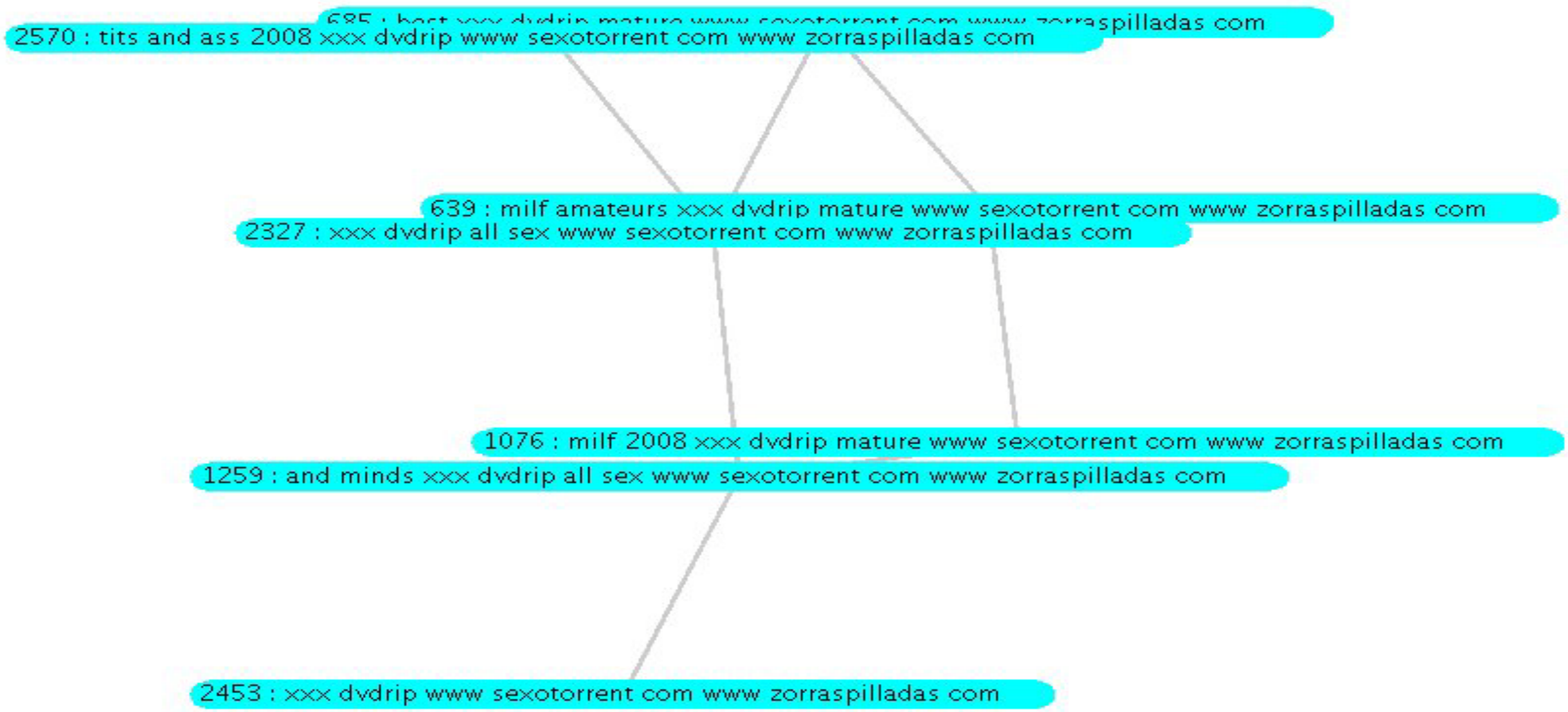


  - Extraction conjointes d'informations sur les thématiques, les fichiers et les utilisateurs

Détection de communautés Structure+Contenu

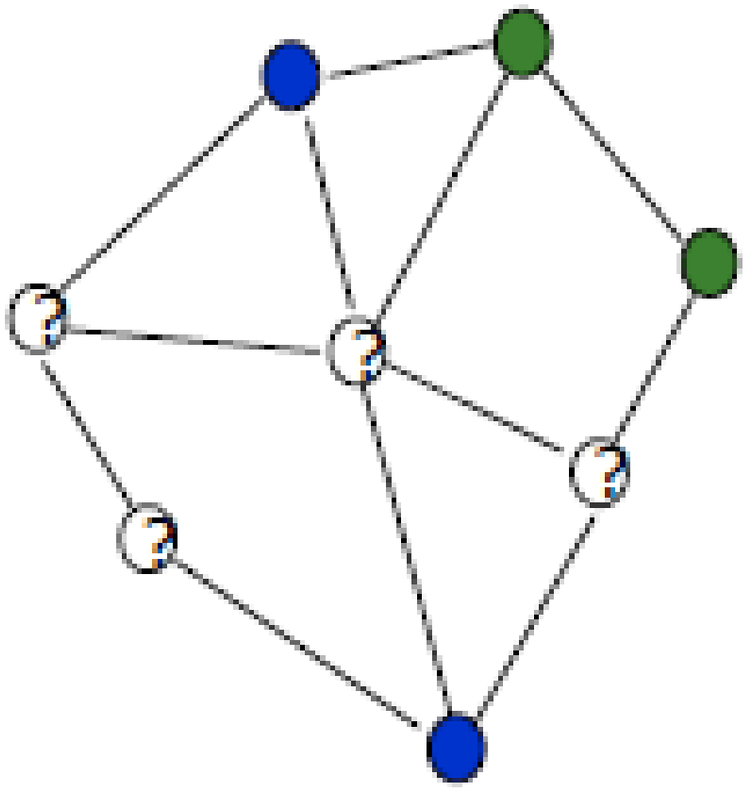


Détection de communautés Structure+Contenu



Détection de communautés Structure+Contenu

Apprentissage supervisé



- Généraliser des étiquettes à un ensemble de nœuds d'un graphe
  - Ex: Vert = pédophilie, Bleu = Pas pédohpilie
  - Identification des utilisateurs/fichiers pédohpiles
- Utilisation de méthodes d'apprentissage semi-supervisée
  - Utilisation du contenu
  - Propagation des labels

Conclusion et pistes futures

- Travaux en cours (P2P) :
  - Détection Communauté d'intérêts
  - Identifier de nouveaux mots clés pédophiles
  - Analyse de la diffusion de fichiers

- Travaux en cours (généraux):
  - Prise en compte de données multi-relationnelles
  - Prise en compte de graphes N-partite

Questions ?