Mesure et analyse des échanges pair-à-pair pour la lutte contre la pédocriminalité

Matthieu Latapy

Matthieu.Latapy@lip6.fr

LIP6 – CNRS et Université Pierre et Marie Curie (UPMC – Paris 6)

Positionnement

Recherche ⟷ **Applications**

problèmes scientifiques — pertinence applicative

être aussi pertinents que possible pour l'application

interactions avec les milieux applicatifs
identification de problématiques
évaluation des résultats
mise en œuvre

PROJET EXPLORATOIRE

résultats, prototypes, méthodes, ...
autres applications (télécoms, politique, ...)

Positionnement

Recherche ⟷ **Applications**

problèmes scientifiques — pertinence applicative

être aussi pertinents que possible pour l'application

interactions avec les milieux applicatifs
identification de problématiques
évaluation des résultats
mise en œuvre

PROJET EXPLORATOIRE

résultats, prototypes, méthodes, ...
autres applications (télécoms, politique, ...)

Positionnement

Recherche ⟷ **Applications**

problèmes scientifiques — pertinence applicative

être aussi pertinents que possible pour l'application

interactions avec les milieux applicatifs
identification de problématiques
évaluation des résultats
mise en œuvre

PROJET EXPLORATOIRE

résultats, prototypes, méthodes, ...
autres applications (télécoms, politique, ...)

Partenaires

- Groupes :
  - **CNRS et UPMC**, France — réseaux, analyse
  - INRIA Lorraine, France — réseaux, analyse
  - UCC, Irlande — psychologie appliquée
  - UL, Slovénie — statistiques, réseaux sociaux
  - FDN, Pologne — association, diffusion
- Financeurs :
  - Communauté Européenne, *Safer Internet Plus*
  - France, *Agence Nationale de la Recherche*

3 ans (2007 – 2010)
> 20 chercheurs impliqués
> 800 KEuros sur trois ans

Partenaires

- Groupes :
  - **CNRS et UPMC**, France — réseaux, analyse
  - INRIA Lorraine, France — réseaux, analyse
  - UCC, Irlande — psychologie appliquée
  - UL, Slovénie — statistiques, réseaux sociaux
  - FDN, Pologne — association, diffusion
- Financeurs :
  - Communauté Européenne, *Safer Internet Plus*
  - France, *Agence Nationale de la Recherche*

3 ans (2007 – 2010)
> 20 chercheurs impliqués
> 800 KEuros sur trois ans

Partenaires

- Groupes :
  - **CNRS et UPMC**, France — réseaux, analyse
  - INRIA Lorraine, France — réseaux, analyse
  - UCC, Irlande — psychologie appliquée
  - UL, Slovénie — statistiques, réseaux sociaux
  - FDN, Pologne — association, diffusion
- Financeurs :
  - Communauté Européenne, *Safer Internet Plus*
  - France, *Agence Nationale de la Recherche*

3 ans (2007 – 2010)
> 20 chercheurs impliqués
> 800 KEuros sur trois ans

Activités

- **mesure**
  - masse d'utilisateurs et d'échanges
  - en continu dans le temps
  - protocoles complexes, peu documentés
  - lois, éthique
- **analyse**
  - information pertinente
  - données massives
  - données complexes, peu structurées

objectifs ambitieux
nombreux défis

Activités

- **mesure**
  - masse d'utilisateurs et d'échanges
  - en continu dans le temps
  - protocoles complexes, peu documentés
  - lois, éthique
- **analyse**
  - information pertinente
  - données massives
  - données complexes, peu structurées

objectifs ambitieux
nombreux défis

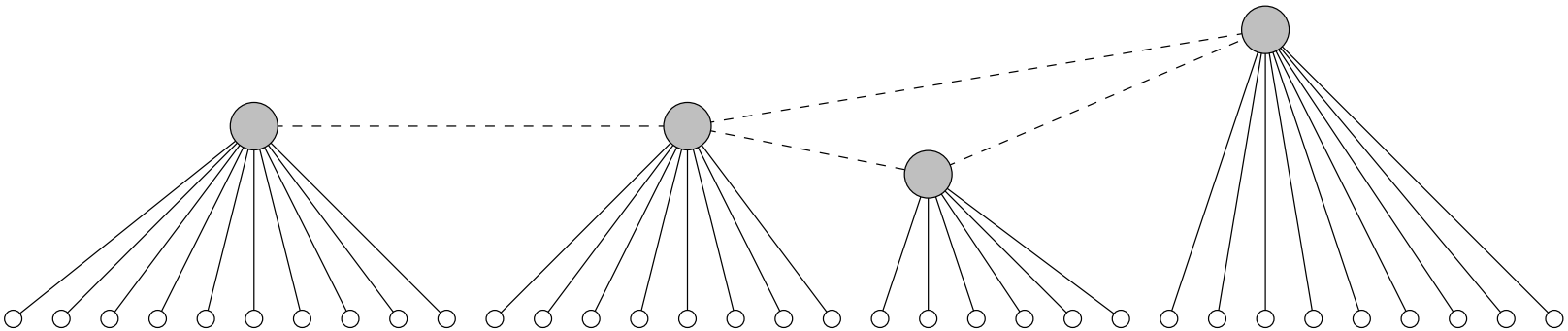
Activités

- **mesure**
  - masse d'utilisateurs et d'échanges
  - en continu dans le temps
  - protocoles complexes, peu documentés
  - lois, éthique
- **analyse**
  - information pertinente
  - données massives
  - données complexes, peu structurées

objectifs ambitieux
nombreux défis

Mesure d'*eDonkey*

Principe de fonctionnement



serveurs + clients

Mesure sur serveur

- **10 semaines en continu**
- **1 milliard de messages**
- **89 millions de pairs (IP)**
- **275 millions de fichiers (hash code)**
- **24 millions de noms de fichiers distincts**
- **116 millions de recherches par mots-clés distinctes**
- **6,6 millions de mots distincts, 1,2 millions apparaissant > 100 fois**

Mesure par client

Envoi périodique de requêtes choisies.

Mesures à la gendarmerie de Bordeaux :

- 8 mots clés spécifiques, 3 jours (terminé)
- 8 mots clés spécifques, 7 généralistes, 1 mois (en cours)

- 120 000 fichiers distincts
- 50 000 clients distincts
- 3 000 clients en France sur 12 000 localisés

mesures distribuées
longues durée
choix de mots-clés/fichiers ?

Mesure par client

Envoi périodique de requêtes choisies.

Mesures à la gendarmerie de Bordeaux :

- 8 mots clés spécifiques, 3 jours (terminé)
- 8 mots clés spécifques, 7 généralistes, 1 mois (en cours)

- 120 000 fichiers distincts
- 50 000 clients distincts
- 3 000 clients en France sur 12 000 localisés

mesures distribuées
longues durée
choix de mots-clés/fichiers ?

Mesure par client

Envoi périodique de requêtes choisies.

Mesures à la gendarmerie de Bordeaux :

- 8 mots clés spécifiques, 3 jours (terminé)
- 8 mots clés spécifques, 7 généralistes, 1 mois (en cours)

- 120 000 fichiers distincts
- 50 000 clients distincts
- 3 000 clients en France sur 12 000 localisés

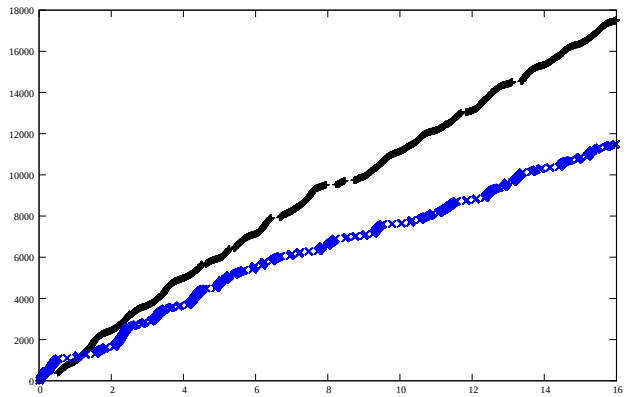
mesures distribuées
longues durée
choix de mots-clés/fichiers ?

Mesure par honeypot

Déclaration de fichiers, attente de requêtes.

Mesures préliminaires :

- un mois en continu
- distribué sur 40 machines



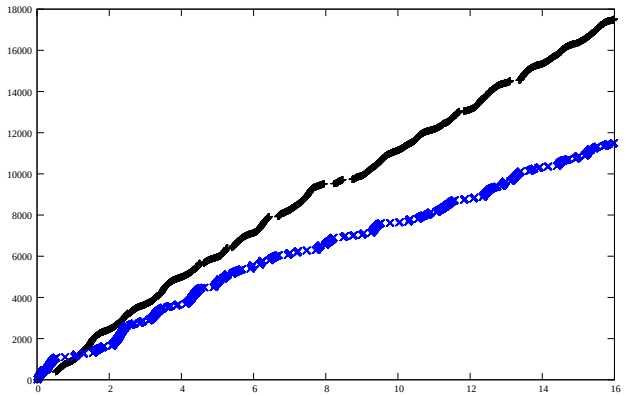
mesures distribuées
longues durée
choix de mots-clés/fichiers ?

Mesure par honeypot

Déclaration de fichiers, attente de requêtes.

Mesures préliminaires :

- un mois en continu
- distribué sur 40 machines



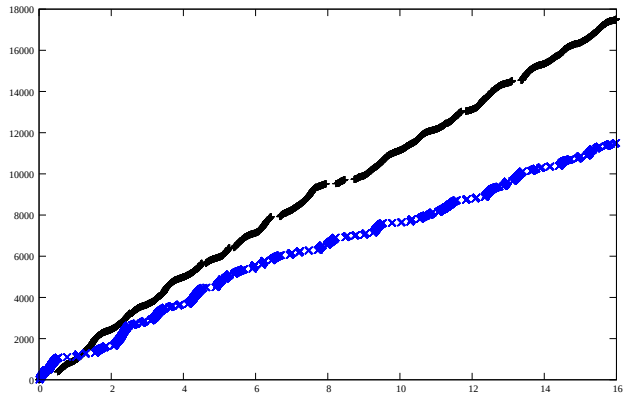
mesures distribuées
longues durée
choix de mots-clés/fichiers ?

Mesure par honeypot

Déclaration de fichiers, attente de requêtes.

Mesures préliminaires :

- un mois en continu
- distribué sur 40 machines



mesures distribuées
longues durée
choix de mots-clés/fichiers ?

Données au format XML

```
<opcode dir="received" TS="284400.777619"
 IP="0002962857" type="high" port="1047">
 <OP_GLOBSEARCHREQ>
  <tags count="3">
   <bool>0</bool>
   <anon-string>d6eebccdd10bc7af0fd54b2bde09f745</
   <named-tag>
    <name-type>byte</name-type>
    <name-value>3</name-value>
    <name-meaning>FILETYPE</name-meaning>
    <string>Audio</string>
   </named-tag>
  </tags>
 </OP_GLOBSEARCHREQ>
</opcode>
```

Interface web

Navigation entre fichiers et pairs.

Données pré-traitées :

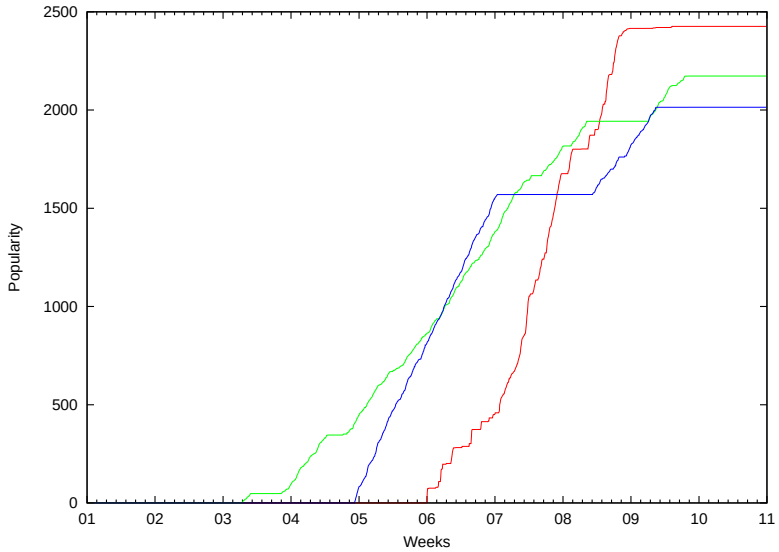
- Qui demande/fournit un fichier donné ?
- Quels fichiers demande/fournit un pair donné ?
- Date de première observation d'un fichier ou d'un pair ?
- Quels noms pour un même fichier ?
- Quelles requêtes pour un même pair ?
- *Content rating* et *fake detection*.
- ...

Analyse

Analyses simples :

- arrivée et diffusion d'un fichier
- proportion de contenu pédo vs non pédo
- âges (noms de fichiers et requêtes)
- mots-clés pédo
- *content rating* et *fake detection*
- ...

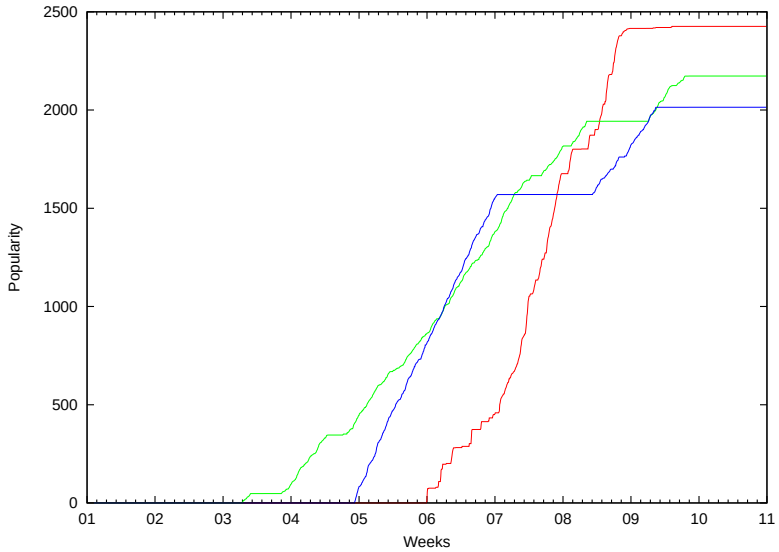
Arrivée et diffusion



x : temps (semaines)
y : popularité du fichier (= nb de personnes le recherchant)

arrivée et départ
types de fichiers ?

Arrivée et diffusion



x : temps (semaines)

y : popularité du fichier (= nb de personnes le recherchant)

arrivée et départ
types de fichiers ?

Proportion d'activité pédo

type	fichiers	requêtes
pédo	0.11 %	0.13 %
madonna	0.15 %	0.06 %
porn	0.8 %	0.05 %

activité pédo sur-représentée dans les requêtes

demande $>$ offre

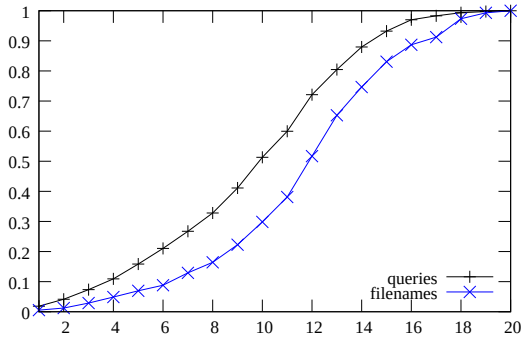
Proportion d'activité pédo

type	fichiers	requêtes
pédo	0.11 %	0.13 %
madonna	0.15 %	0.06 %
porn	0.8 %	0.05 %

activité pédo sur-représentée dans les requêtes

demande $>$ offre

Analyse – âges



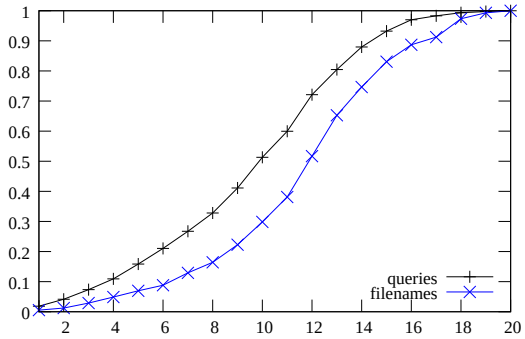
x : âge sous la forme *xyo*

y : nombre d'apparitions avec un âge $\leq x$

$\leq$ 10 ans : 40% (requêtes) et 50% (fichiers)

$\leq$ 5 ans : 15% (requêtes) et 7% (fichiers)

Analyse – âges



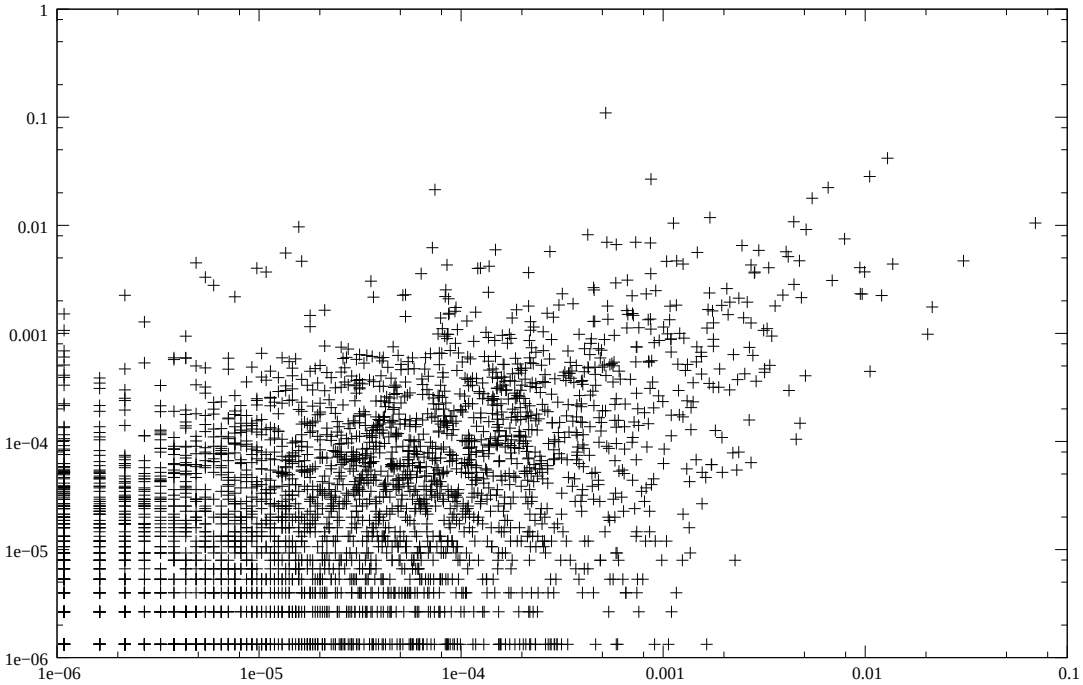
x : âge sous la forme *xyo*

y : nombre d'apparitions avec un âge $\leq x$

$\leq$ 10 ans : 40% (requêtes) et 50% (fichiers)

$\leq$ 5 ans : 15% (requêtes) et 7% (fichiers)

Mots-clés

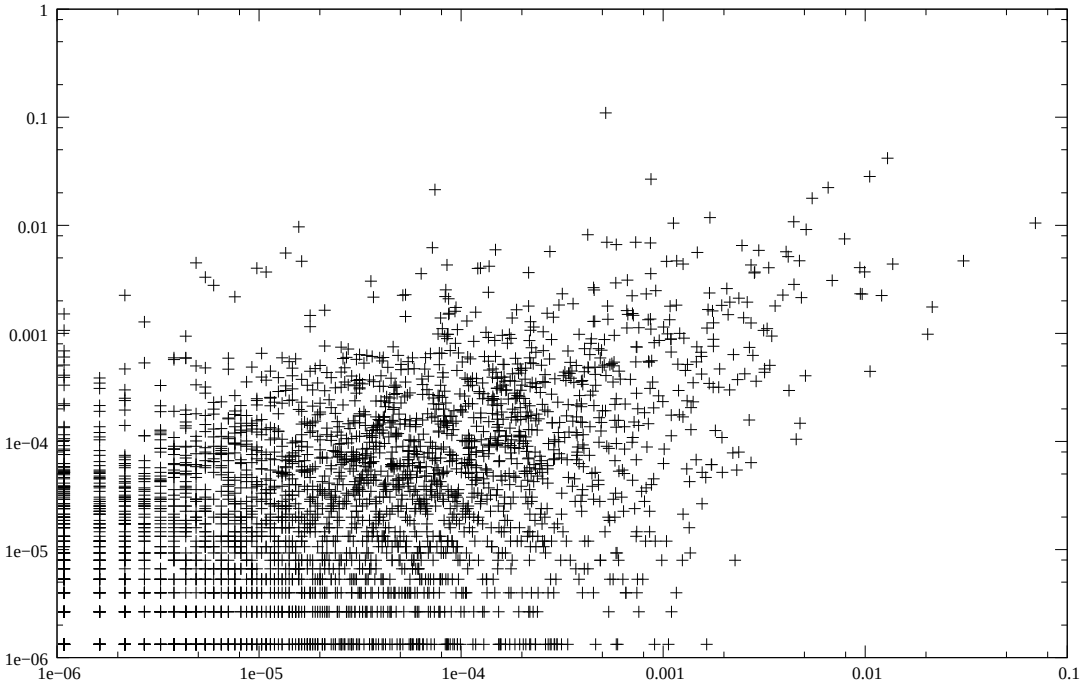


frequences comparees

x : fréquence d'apparition avec mots-clés pédo
y : fréquence d'apparition avec *porn*

découverte de nouveaux mots-clés pédo ?

Mots-clés



frequences comparees

x : fréquence d'apparition avec mots-clés pédo
y : fréquence d'apparition avec *porn*

découverte de nouveaux mots-clés pédo ?

Content rating et *fake detection*

Répondre automatiquement à :

- tel ficher a-t-il un contenu à caractère pornographique ? pédophile ?
- a-t-il un contenu significativement différent de son nom ?

pour aider à la classification et protéger les utilisateurs.

Actuellement : basé sur les mots-clés seulement.

résultats mitigés

Content rating et *fake detection*

Répondre automatiquement à :

- tel ficher a-t-il un contenu à caractère pornographique ? pédophile ?
- a-t-il un contenu significativement différent de son nom ?

pour aider à la classification et protéger les utilisateurs.

Actuellement : basé sur les mots-clés seulement.

résultats mitigés

Content rating et *fake detection*

Répondre automatiquement à :

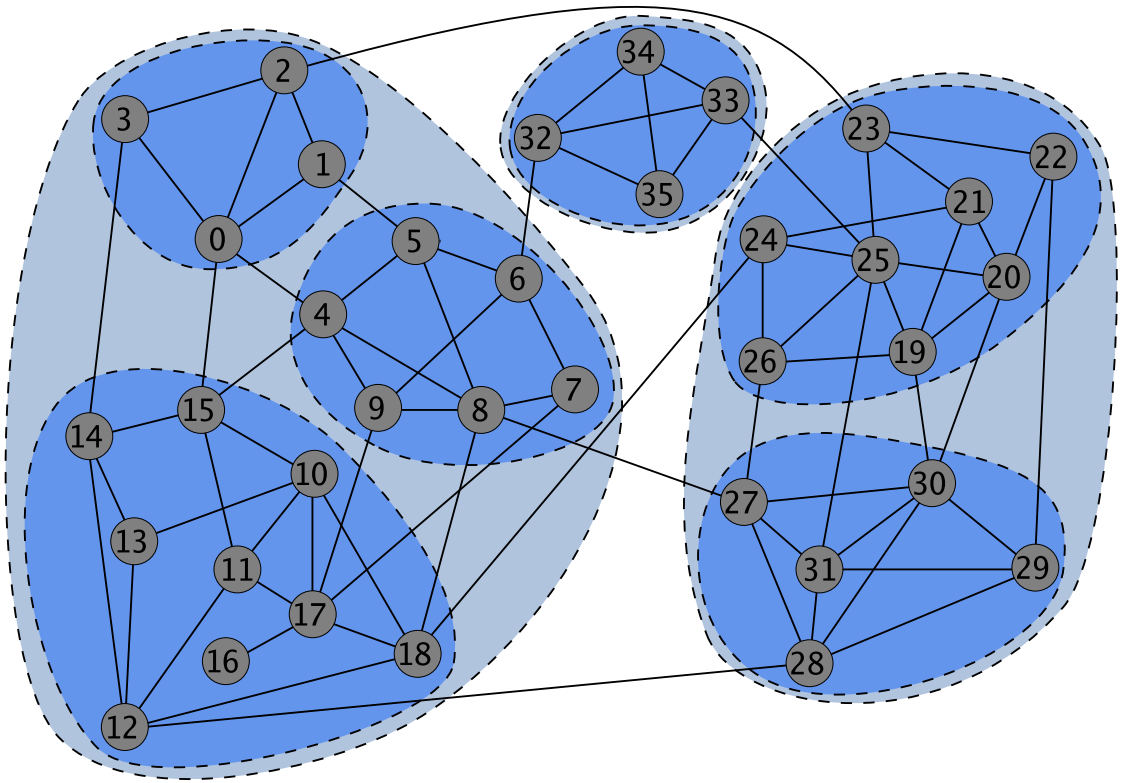
- tel ficher a-t-il un contenu à caractère pornographique ? pédophile ?
- a-t-il un contenu significativement différent de son nom ?

pour aider à la classification et protéger les utilisateurs.

Actuellement : basé sur les mots-clés seulement.

résultats mitigés

Analyse – graphes



deux fichiers sont "reliés"
si
beaucoup de pairs les fournissent tous deux

Conclusion

- **Le projet :**
  - Données massives
  - Analyses poussées
  - Exploratoire
- **Premiers résultats :**
  - Mesures sur serveur
  - Analyses simples
  - Autres mesures/analyses en cours
- **Nos attentes :**
  - Retour sur les résultats
  - Questions prioritaires
  - Expertise (mots-clés, fichiers intéressants, ...)