

Estimation fiable de propriétés dans des systèmes dynamiques : la durée de session dans les systèmes P2P

Lamia Benamara, Clémence Magnien

Complex Networks
LIP6 - CNRS - Université Pierre et Marie Curie

16 février 2010

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

- 1 Contexte
- 2 Données et propriété étudiée
- 3 Méthodologie et résultats
- 4 Conclusion et perspectives

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

- 1 Contexte
- 2 Données et propriété étudiée
- 3 Méthodologie et résultats
- 4 Conclusion et perspectives

informatique : web, internet, pair-à-pair, usages, etc

sciences sociales : amitié, communications, collaboration, échanges, économie, etc

biologie : cerveau, gènes, protéines, écosystèmes, etc

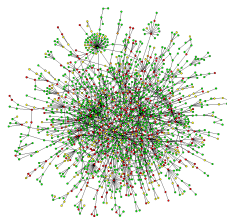
linguistique : synonymie, co-occurrence, etc

transport : routier, aérien, électrique, etc

etc, etc

réseaux de relations

contextes très différents



**la plupart des graphes de terrain ont des
propriétés non-triviales en commun [WS98]**

propriétés en commun :

- distance moyenne faible (*petit-monde*)
- degrés hétérogènes (loi de puissance, *scale-free*)
- densité globale faible vs densité locale forte (*clustering*)

Beaucoup de travaux sur l'étude des graphes de terrain
→ ensemble d'outils et de notions reconnus

la plupart des graphes de terrain ont des propriétés non-triviales en commun [WS98]

propriétés en commun :

- distance moyenne faible (*petit-monde*)
- degrés hétérogènes (loi de puissance, *scale-free*)
- densité globale faible vs densité locale forte (*clustering*)

Beaucoup de travaux sur l'étude des graphes de terrain
→ ensemble d'**outils** et de **notions** reconnus

Et la dynamique ?

les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

(Tous) les graphes de terrain sont **dynamiques**

Ajout et suppression au fil du temps

- de nœuds
- de liens

Peu ou pas de notions établies pour l'étude de la dynamique

Et la dynamique ?

les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

(Tous) les graphes de terrain sont **dynamiques**

Ajout et suppression au fil du temps

- de nœuds
- de liens

Peu ou pas de notions établies pour l'étude de la dynamique

Biais dans la mesure de la dynamique

les systèmes
P2P

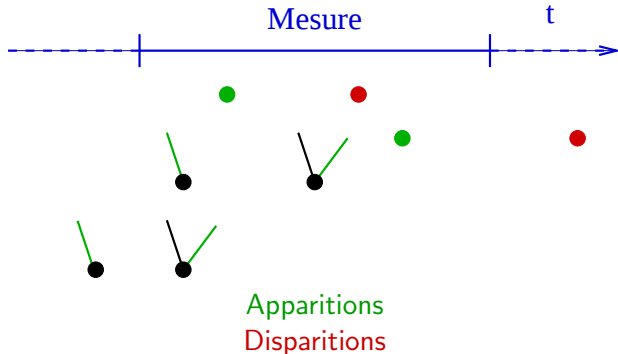
Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

Période de mesure finie $\rightarrow$ événements non observés



Biais dans l'estimation des propriétés dynamiques

- 1 Contexte
- 2 Données et propriété étudiée
- 3 Méthodologie et résultats
- 4 Conclusion et perspectives

Système P2P eDonkey

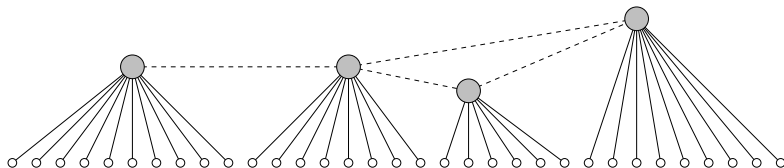
les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives



utilisateur

serveur

mots-clés



liste de fichiers

fichier



fournisseurs

Type de données

- Capture des échanges sur un gros serveur *eDonkey*
- Période de mesure: 10 semaines
- Ensembles de requêtes / réponses, **information temporelle**
- 1 milliard de msg, 89 millions de pairs

Propriété étudiée

Durée des sessions (distribution)

Session : intervalle de temps où l'utilisateur est connecté au système

Type de données

- Capture des échanges sur un gros serveur *eDonkey*
- Période de mesure: 10 semaines
- Ensembles de requêtes / réponses, **information temporelle**
- 1 milliard de msg, 89 millions de pairs

Propriété étudiée

Durée des sessions (distribution)

Session : intervalle de temps où l'utilisateur est connecté au système

Problème : identification des utilisateurs

les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

Une IP $\Rightarrow$ plusieurs utilisateurs

- plusieurs ordinateurs, même domicile
- cybercafés, ...
- plusieurs sessions $\Rightarrow$ une seule session

Un utilisateur $\Rightarrow$ plusieurs IP

- adresses dynamiques
- cybercafés, ...
- un utilisateur $\Rightarrow$ plusieurs utilisateurs
- pas de changement pendant une session

Problème : identification des utilisateurs

les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives

Une IP $\Rightarrow$ plusieurs utilisateurs

- plusieurs ordinateurs, même domicile
- cybercafés, ...
- plusieurs sessions $\Rightarrow$ une seule session

Un utilisateur $\Rightarrow$ plusieurs IP

- adresses dynamiques
- cybercafés, ...
- un utilisateur $\Rightarrow$ plusieurs utilisateurs
- pas de changement pendant une session

Notre choix : un utilisateur = une adresse IP.

Problème : définition de session

les systèmes
P2P

Contexte

Données et
propriété
étudiée

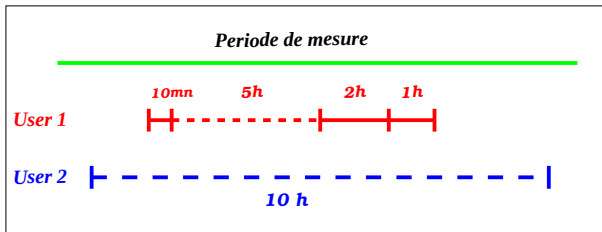
Méthodologie
et résultats

Conclusion et
perspectives

Sessions **son données** par le protocole

→ besoin de l'inférer à partir des **requêtes**

- Seuil choisi: 3 heures



Utilisateur 1:

- session 1 = 10 minutes
- session 2 = 3 heures

Utilisateur 2:

- 2 sessions = 0

- 1 Contexte
- 2 Données et propriété étudiée
- 3 Méthodologie et résultats
- 4 Conclusion et perspectives

Éstimation de la durée des sessions

les systèmes
P2P

Contexte

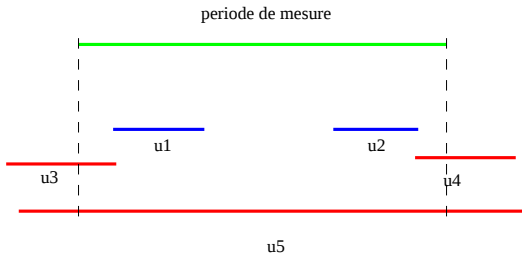
Données et
propriété
étudiée

Méthodologie
et résultats

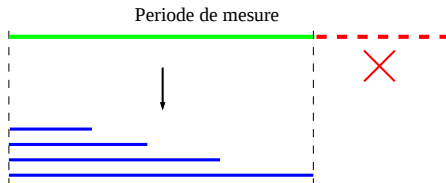
Conclusion et
perspectives

Certaines session:

- commencent avant le début de mesure
 - finissent après la fin de mesure
- ↗ période grande $\implies$ ↗ le biais est petit



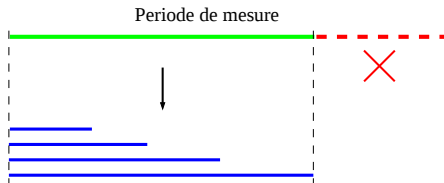
- impossible d'étendre la mesure
- simulation de mesures plus courtes



évolution en fonction de la durée de mesure

- Fluctuations $\Rightarrow$ pas de conclusion
- Stabilité $\Rightarrow$ indépendant de la durée de mesure (a priori).

- impossible d'étendre la mesure
- simulation de mesures plus courtes



évolution en fonction de la durée de mesure

- Fluctuations $\Rightarrow$ pas de conclusion
- Stabilité $\Rightarrow$ indépendant de la durée de mesure (a priori).

Résultats

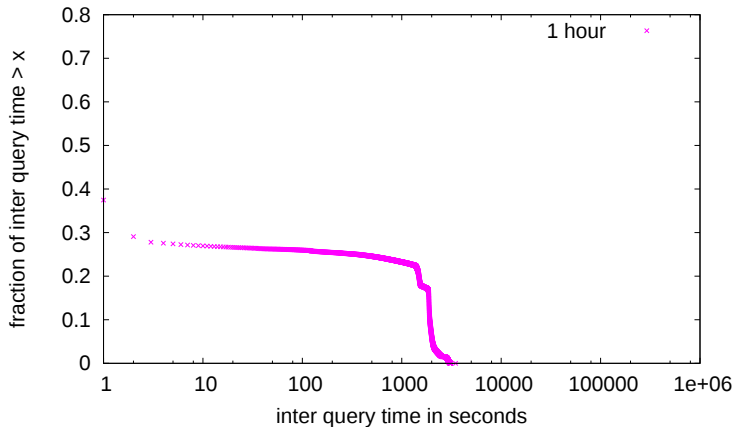
les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives



Résultats

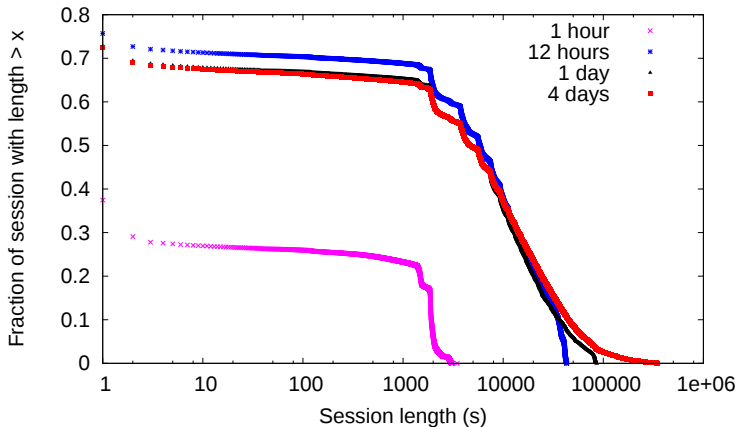
les systèmes
P2P

Contexte

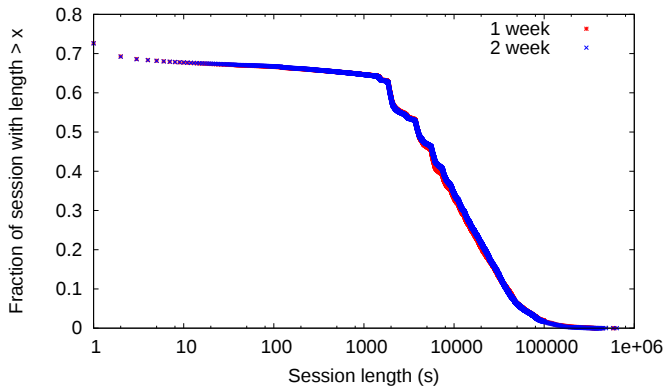
Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives



- Évolution
- Changement de forme



- Convergence
- ≥ 1 semaine $\Rightarrow$ Suffisant

Résultats

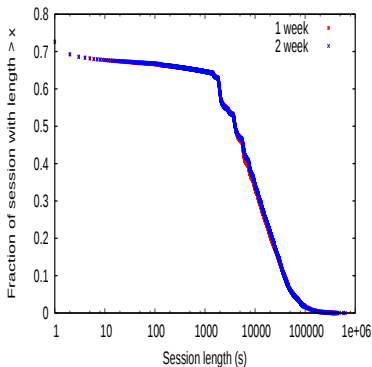
les systèmes
P2P

Contexte

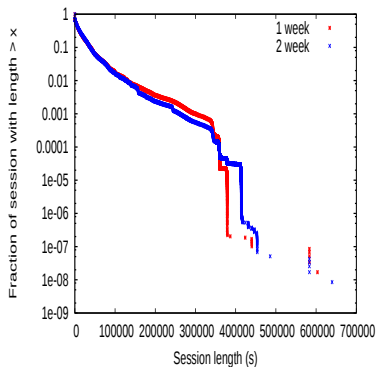
Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives



échelle log lin



échelle lin log

- 99% identiques
- Valeurs extrêmes

Moyenne et écart type

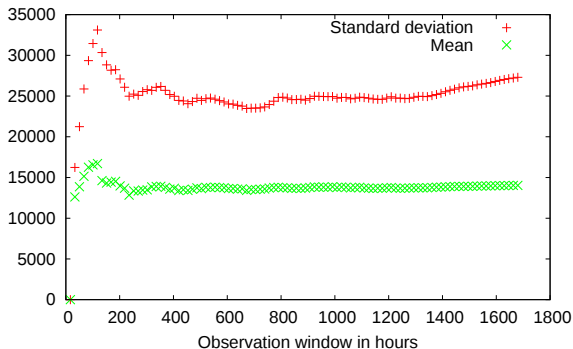
les systèmes
P2P

Contexte

Données et
propriété
étudiée

Méthodologie
et résultats

Conclusion et
perspectives



1 semaine = 168 h

Moyenne : stabilisation → caractérisation de la partie normale

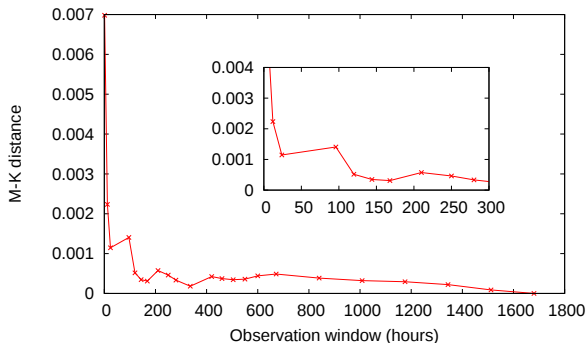
Écart type : non stabilisation → valeurs extrêmes

Distance Monge-Kantorovich

- $(\sum_k |p(k) - p'(k)|) / k_{\max}$.

Distance Monge-Kantorovich

- $(\sum_k |p(k) - p'(k)|) / k_{\max}$.



Convergence après ~ une semaine

- 1 Contexte
- 2 Données et propriété étudiée
- 3 Méthodologie et résultats
- 4 Conclusion et perspectives

- Introduit une méthodologie
- Durée de mesure nécessaire ~ 1 semaine.
- Forme de distribution :
 - Comportement normal
 - Valeurs extrêmes $\rightarrow$ impossible à caractériser
- Méthodologie générique $\implies$ s'applique à d'autres propriétés et systèmes

- Introduit une méthodologie
- Durée de mesure nécessaire ~ 1 semaine.
- Forme de distribution :
 - Comportement normal
 - Valeurs extrêmes $\rightarrow$ impossible à caractériser
- Méthodologie générique $\implies$ s'applique à d'autres propriétés et systèmes

- Autres définitions d'utilisateurs et de sessions
- Méthode de séparation des deux classes de comportement (normal et extrême)
- Sessions de longueur 0
- Simulations

Brique de base pour étudier

- d'autres propriétés
- d'autres systèmes