

Intégrer mesure, métrologie et analyse pour l'étude des graphes de terrain dynamiques

Clémence Magnien

LIP6 - CNRS - Université Pierre et Marie Curie

6 juillet 2010

Habilitation à diriger des recherches



Plan

- 1 Contexte
- 2 Mesure
- 3 Métrologie
- 4 Analyse
- 5 Conclusion

Graphes de terrain

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

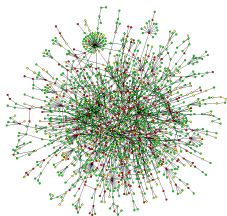
sciences sociales : amitié, communications, collaboration, échanges, économie, etc

biologie : cerveau, gènes, protéines, écosystèmes, etc

linguistique : synonymie, co-occurrence, etc

transport : routier, aérien, électrique, etc

etc, etc



Cohérence d'ensemble

Ressemblance structurelle

Propriétés communes

Problématiques communes

- **Mesure** : Graphes non donnés *a priori*
→ Acquérir l'information (partielle)
- **Métrologie** : Information partielle → biais statistique
→ Étude/correction du biais, mesure sans biais
- **Analyse** : Objets de grande taille
→ Décrire (propriétés statistiques/structurelles)

Cohérence d'ensemble

Ressemblance structurelle

Propriétés communes

Problématiques communes

- **Mesure** : Graphes non donnés *a priori*
→ Acquérir l'information (partielle)
- **Métrologie** : Information partielle → biais statistique
→ Étude/correction du biais, mesure sans biais
- **Analyse** : Objets de grande taille
→ Décrire (propriétés statistiques/structurelles)

Dynamique

La plupart des graphes de terrain sont **dynamiques**

Ajouts et suppressions

- de nœuds
- de liens

Récemment, quelques travaux sur la dynamique

- Cas particuliers
- Questions spécifiques

Manque de méthodes **générales**

Plan

- 1 Contexte
- 2 Mesure**
- 3 Métrologie
- 4 Analyse
- 5 Conclusion

Mesure

But

Acquérir l'information sur les nœuds et liens et leur évolution

Contributions

Système *eDonkey*

OA

Mesure par client, serveur, honeypots

[HotP2P'09] ×2

Un radar pour l'internet

FO

Vue ego-centrée, nouvelle approche

[ADN'08]

Vue ego-centrée

Topologie de l'internet

Routeurs, liens au niveau IP

Mesure : outil traceroute

- Information partielle (et biaisée)
- Coûteux → mesure périodique : fréquence faible

Approche orthogonale : vue ego-centrée

- | | |
|--------------------------|----------------------------|
| ● un moniteur | ● objet bien défini |
| ● plusieurs destinations | ● fréquence élevée : radar |

Vue ego-centrée

Topologie de l'internet

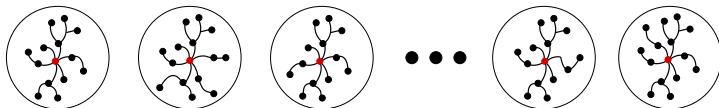
Routeurs, liens au niveau IP

Mesure : outil traceroute

- Information partielle (et biaisée)
- Coûteux → mesure périodique : fréquence faible

Approche orthogonale : vue ego-centrée

- | | |
|--------------------------|-----------------------------------|
| • un moniteur | • objet bien défini |
| • plusieurs destinations | • fréquence élevée : radar |



Outil et paramètres

Paramètres

- fréquence
- nombre de destinations
- distance max
- ...

→ tester **toutes** les combinaisons ?

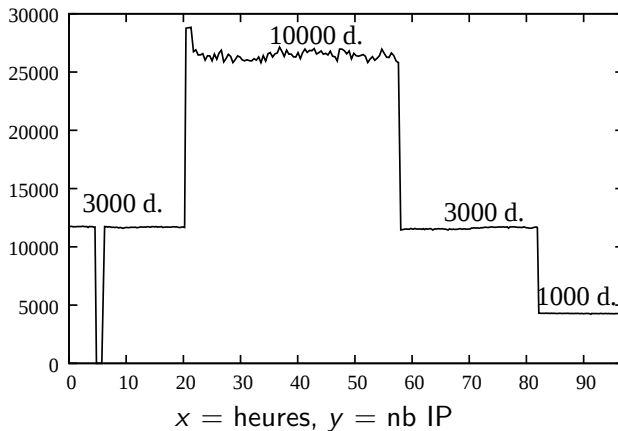
Notre approche

Paramètres de base

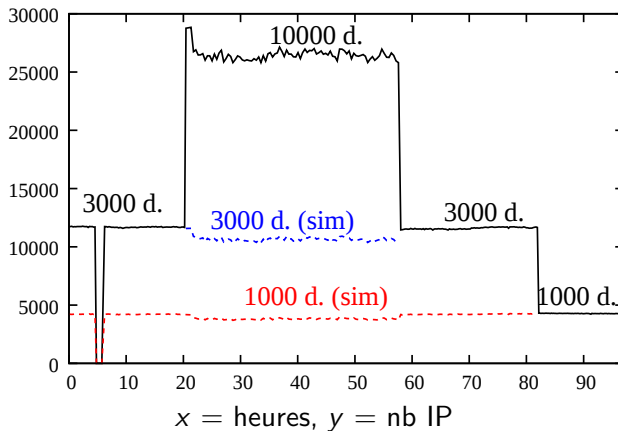
Changer un paramètre à la fois

Moniteur de contrôle

Nombre de destinations



Nombre de destinations



Mesure – Conclusion

Mesures

- une centaine de moniteurs
- plusieurs mois

Outil et données publiquement disponibles

- Importance de tester les paramètres
→ techniques générales ?
- Mesures publiquement disponibles
→ autres travaux, validation des analyses
- Paramètres documentés
→ reproductibilité, interprétation

Plan

- 1 Contexte
- 2 Mesure
- 3 Métrologie**
- 4 Analyse
- 5 Conclusion

Métrologie

But

Mesure $\rightarrow$ biais dans les propriétés observées
Étudier/corriger ce biais

Contributions

Load-balancing

Paris traceroute [IMC'06], [ComNet'08]

Dynamique de la mesure

Évolution de l'échantillon $\rightarrow$ représentatif ? [Infocom'08]

Mesure finie $\rightarrow$ biais LB

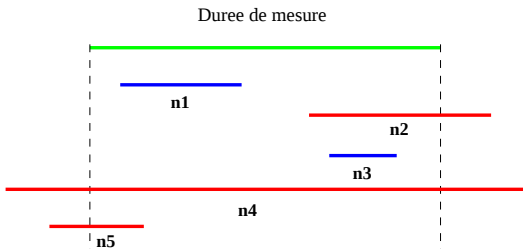
Méthodologie pour supprimer ce biais [NetSciCom'10]

Durée de mesure finie

Certains nœuds :

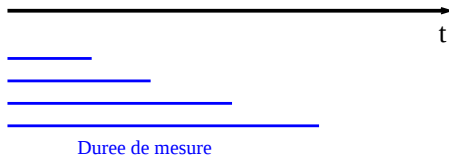
- apparaissent avant le début de mesure
- disparaissent après la fin de mesure

durée ↗ ⇒ biais ↘



Principe

Propriété : évolution en fonction de la durée de mesure



- Stabilité $\implies$ indépendant de la durée de mesure (a priori)
- Fluctuations $\implies$ observation non fiable

Application : durée des sessions

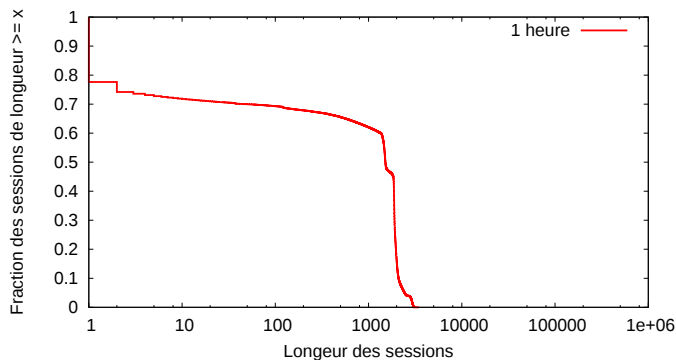
Sessions dans le système P2P *eDonkey*

Temps pendant lequel un utilisateur est connecté au système

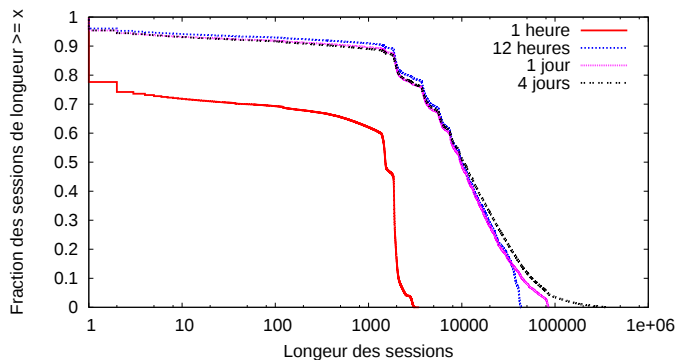
Données

- Capture sur un serveur
- 10 semaines

Résultats

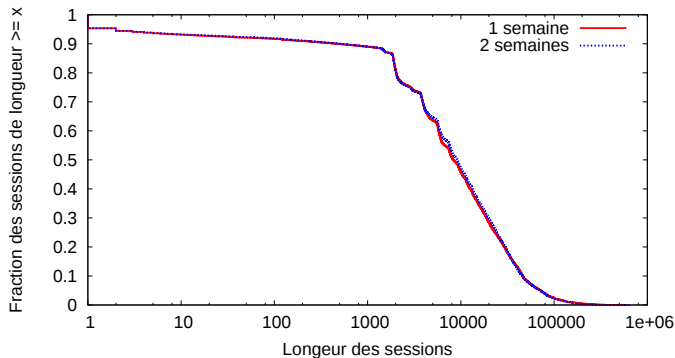


Résultats

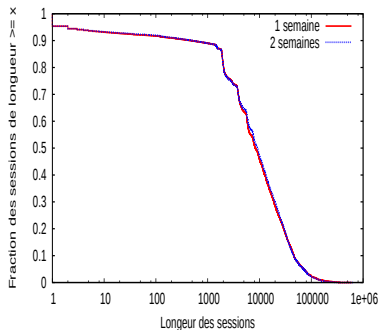


- Évolution
- Changement de forme

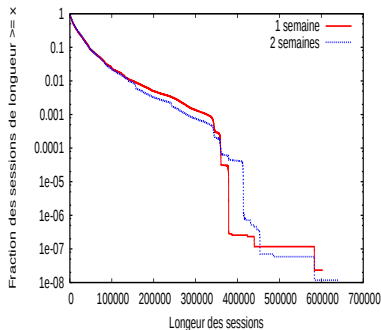
Résultats



- Convergence
- ≥ 1 semaine $\Rightarrow$ suffisant



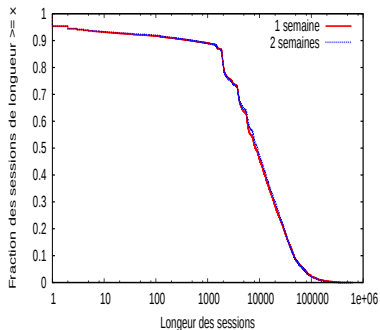
échelle log lin



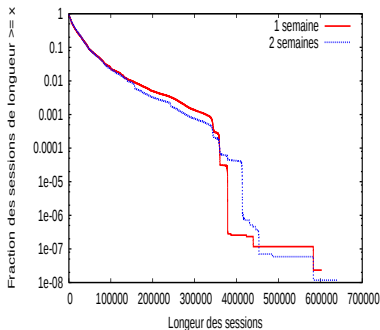
échelle lin log

- 99% identiques
- Valeurs extrêmes

Confirmé par des tests statistiques



échelle log lin



échelle lin log

- 99% identiques
- Valeurs extrêmes

Confirmé par des **tests statistiques**

Métrologie – Conclusion

Introduit une méthodologie

Caractériser une propriété

→ propriétés stationnaires/non stationnaires

Méthodologie **générale** (propriétés, graphes)

→ métrologie $\neq$ mesure

Plan

- 1 Contexte
- 2 Mesure
- 3 Métrologie
- 4 Analyse**
- 5 Conclusion

Analyse

But

Décrire l'évolution du graphe

Contributions

Robustesse

Suppression de nœuds et de liens (pannes, attaques)

[OPODIS'04], [ComSur'10]

Vues ego-centrées de l'internet

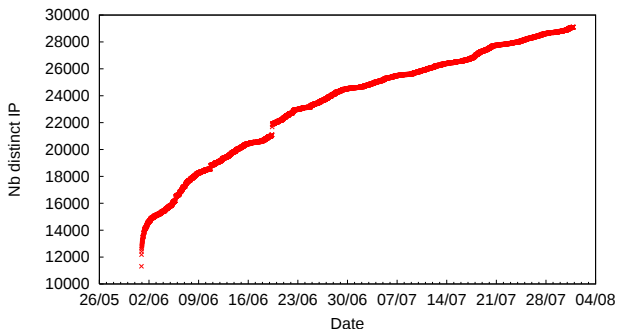
FO, GV

Dynamique des vues ego-centrées

[ICIMP'09]

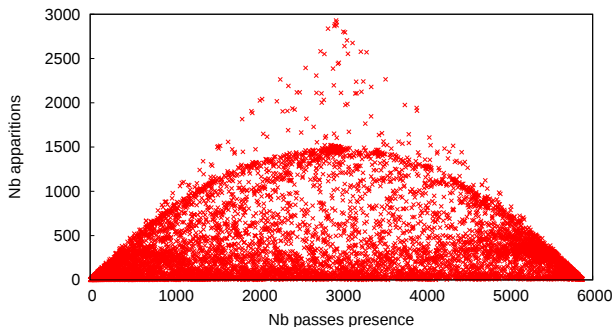
Dynamique : stabilisation ?

Nombre d'IP distinctes observées depuis le début



nouvelles IP observées en permanence
importance des changements de routage

Présences vs apparitions

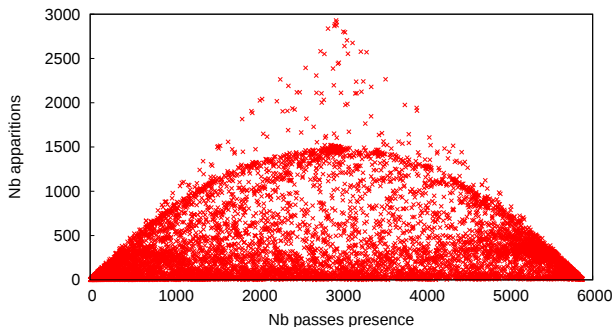


1 2 3 4 5 6 7 8



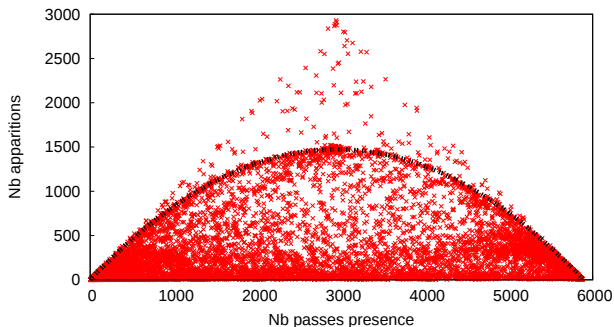
5 présences, 2 apparitions

Présences vs apparitions



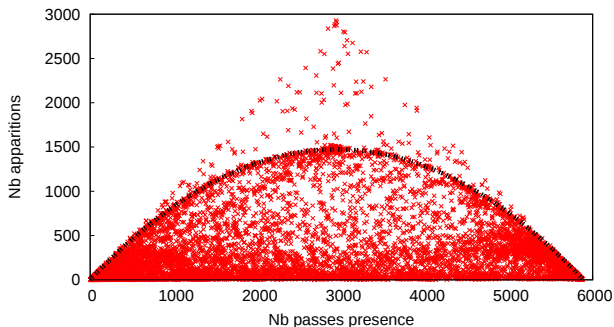
Triangle : $\text{Nb apparitions} \leq \max(\text{nb présences}, \text{nb absences})$

Présences vs apparitions



Parabole : Nombre d'apparitions attendues en moyenne pour ce nombre de présences

Présences vs apparitions



Trois classes d'IP

aléatoires || majorité de stables || clignotent

Analyse – Conclusion

Dynamique des vues ego-centrées

- Mise en évidence une dynamique forte
- Isolation de plusieurs classes d'adresses IP

→ traiter des propriétés plus complexes

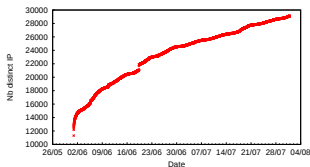
Définir des notions génériques

- Nouvelles statistiques
- Traiter plusieurs cas

Plan

- 1 Contexte
- 2 Mesure
- 3 Métrologie
- 4 Analyse
- 5 Conclusion**

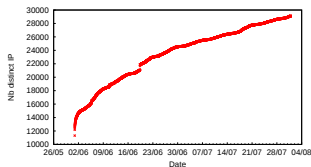
Pour aller plus loin



Évolution des adresses causée par

- le *load-balancing*
 - plus de passes → plus d'IP
- des changements de routage
 - plus de temps → plus d'IP

Pour aller plus loin



Évolution des adresses causée par

- le *load-balancing*
 - plus de passes → plus d'IP
- des changements de routage
 - plus de temps → plus d'IP

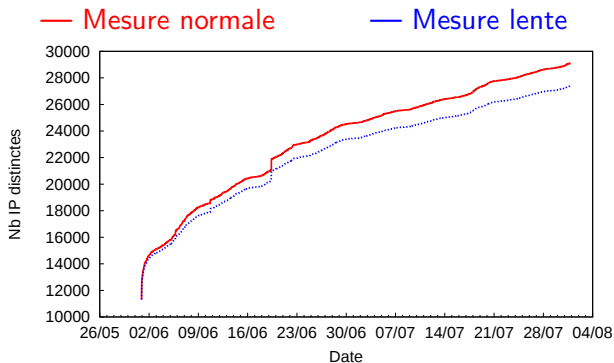
Importance du temps et du nombre de passes : fréquence

Mesure normale et mesure lente

Deux mesures

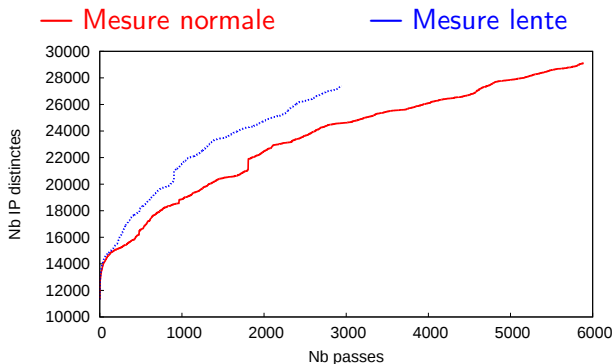
- fréquences différentes
 - même moniteur/destinations
 - même moment
-
- Mesure radar normale
 - → une passe sur deux : mesure lente

Mesure normale et mesure lente



Temps égal, plus de passes : plus d'IP → *load-balancing*

Mesure normale et mesure lente



Nombre de passes égal, plus de temps : plus d'IP
→ changements de routage

Conséquences

Mise en évidence de deux facteurs : **Analyse**

Vitesse d'évolution

Pente de la courbe :

- Dépend **intrinsèquement** de la fréquence de **mesure**
- Pas corrigeable par la **métrologie**

Étudier la pente de la courbe

→ combiner mesure, métrologie et analyse

Conclusion

- Mesure
- Métrologie
- Analyse

Certaines questions **sont au-delà** de ces catégories

→ Autres grandes problématiques ? (**fréquence** de mesure)

Perspectives

Influence fréquence / durée mesure

- Traiter d'autres cas
- Simulations

Cartographie de l'internet / graphes dynamiques

- Compromis exhaustivité / actualité
- Indices de confiance dans les nœuds / liens ?

Outils d'analyse

- Outils et notions standards

Modèles / formalismes