Structure et Dynamique des Réseaux

Métrologie de la topologie de l'internet

Clémence Magnien, Lionel Tabourier, Fabien Tarissan

LIP6 – CNRS and Université Pierre et Marie Curie

prenom.nom@lip6.fr

Outline

Topologie de l'internet

Exploration : `traceroute`
Quelques sources, bcp de destinations :

- On sait qu'on ne voit pas tout
- Vue représentative ? ($\rightarrow$ biais ?)

Distribution des degrés

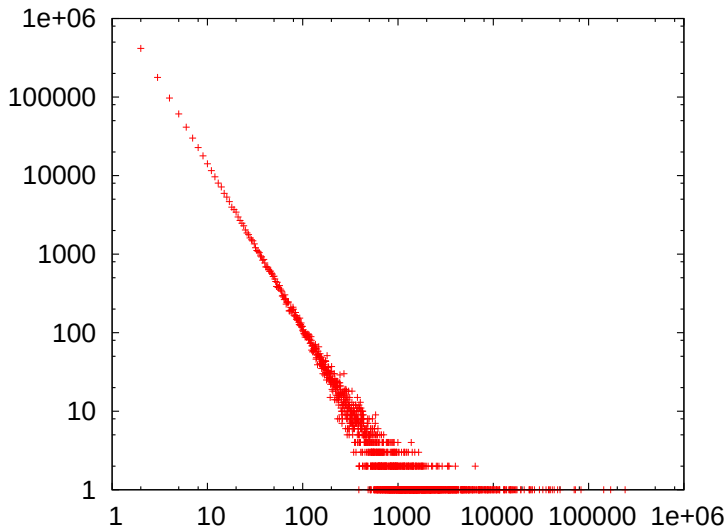
Une propriété dont on a beaucoup parlé :
Distribution des degrés de l'internet en loi de puissance

[Pansiot, Grad, 1998
Faloutsos, Faloutsos, Faloutsos, 1999]

Loi de puissance (power-law)

Loi de puissance

- $N_k \sim k^{-\alpha}$
- droite en échelle log-log



Distributions normalisées

Distribution des degrés, deux choix :

- N_k : nombre de sommets de degré k
- p_k : fraction de sommets de degré k

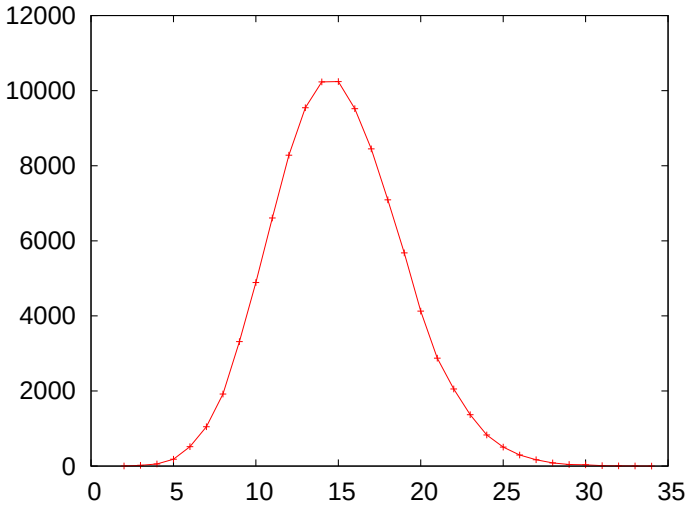
$\rightarrow$ Distribution normalisée

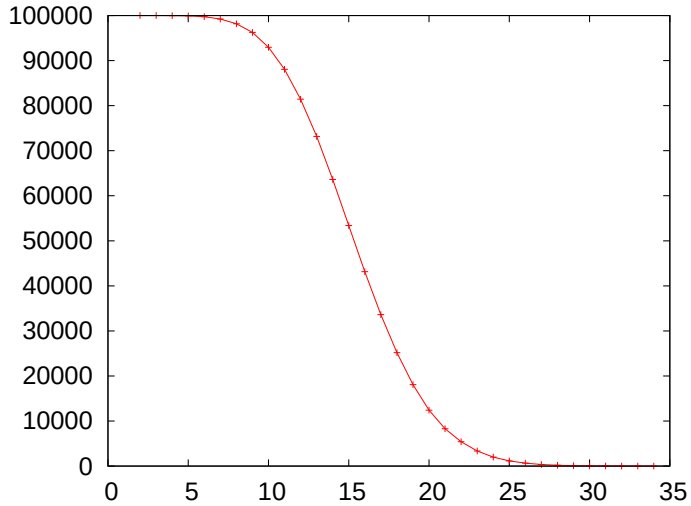
$$p_k = \frac{N_k}{n}$$

Permet de comparer des graphes de tailles différentes

Distribution cumulative inverse

- N_k : nombre de nœuds de degré égal à k
- C_k : nombre de nœuds de degré inférieur ou égal à k

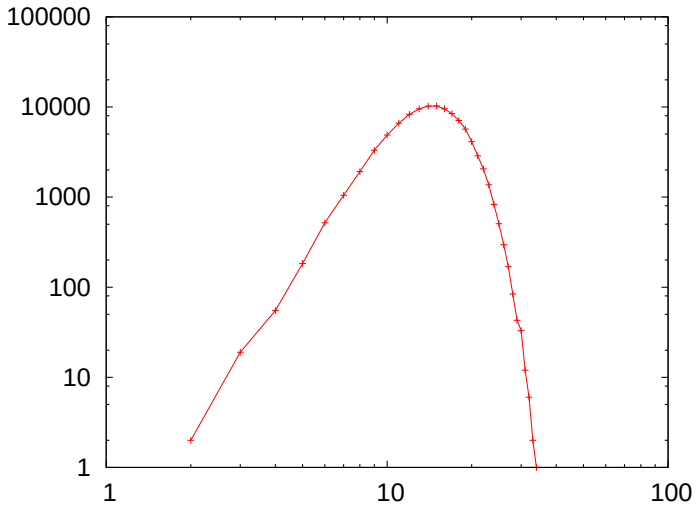


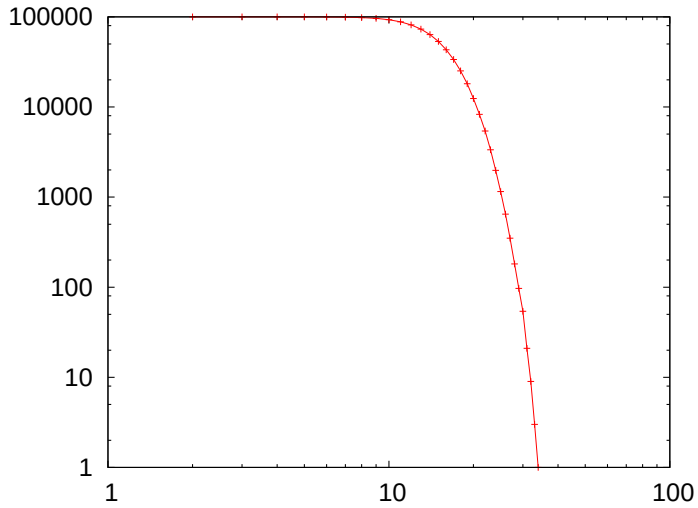


Échelle linéaire

Distribution cumulative inverse

- N_k : nombre de nœuds de degré égal à k
- C_k : nombre de nœuds de degré inférieur ou égal à k





Échelle log-log

Distribution cumulative inverse

- N_k : nombre de nœuds de degré égal à k
- C_k : nombre de nœuds de degré inférieur ou égal à k

Échelle

Distributions homogènes/hétérogènes

- Se distingue sur la distribution normale ou cumulative

Ex : loi de puissance

- $N_k \sim k^{-\alpha} \Longrightarrow C_k \sim k^{-\alpha+1}$

Distribution des degrés observée surprenante $\rightarrow$ biais ?

Conséquences :

Diminuer le biais

Mesure depuis un plus grand nombre de sources

Estimer le biais

Études théoriques et expérimentales

Outline

Outline

Multiplier les sources et les destinations

Intérêt d'utiliser plusieurs sources et destinations

→ Quantité d'information ?
→ Diminution du biais ?

Quantité d'information

[On the Marginal Utility of Network Topology Measurements
Barford, Bestavros, Byers, Crovella, 2001]

Idée

- Utiliser des données issues de mesures (vs simulation)
- Estimer le nombre de nœuds/liens vus en fonction du nombre de sources/destinations

Données

Deux jeux de données

- 8 sources
- 1277 destinations
- 1 traceroute toute les 30 minutes
- 7 mois (?)

- 12 sources
- $>$ 300 000 destinations
- même méthode de mesure
- durée ?

Données

Note :

Intérêt de répéter les mesures ?

Load-balancing, . . .
$\rightarrow$ répéter donne plus d'informations

Plus de détails dans le prochain cours

Méthodologie

Estimer le nombre de nœuds vus en fonction :

- du nombre de sources
- du nombre de destinations

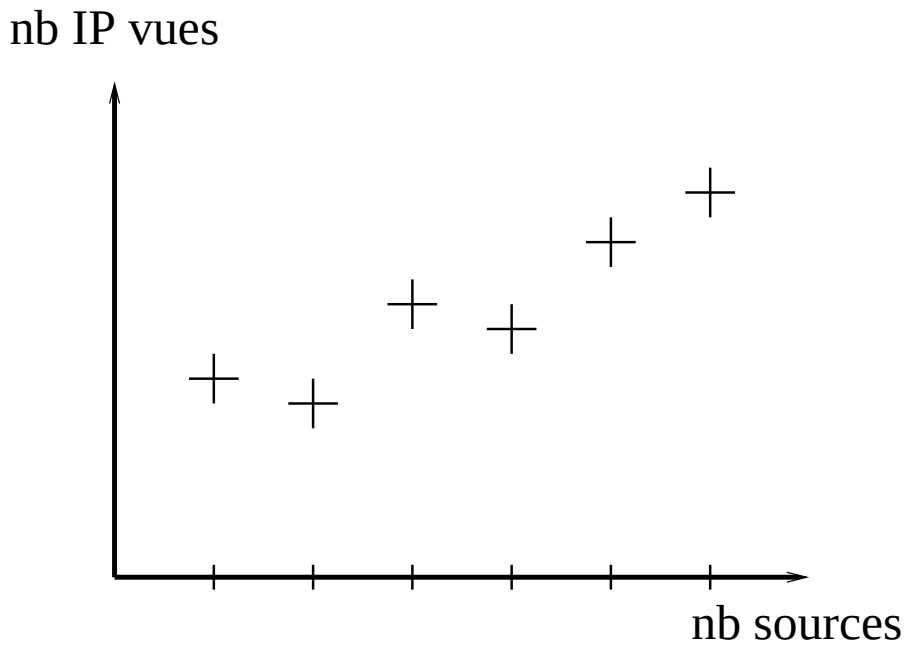
s sources, d destinations
$\rightarrow s \times d$ valeurs possibles

Beaucoup d'informations

Interprétation ?

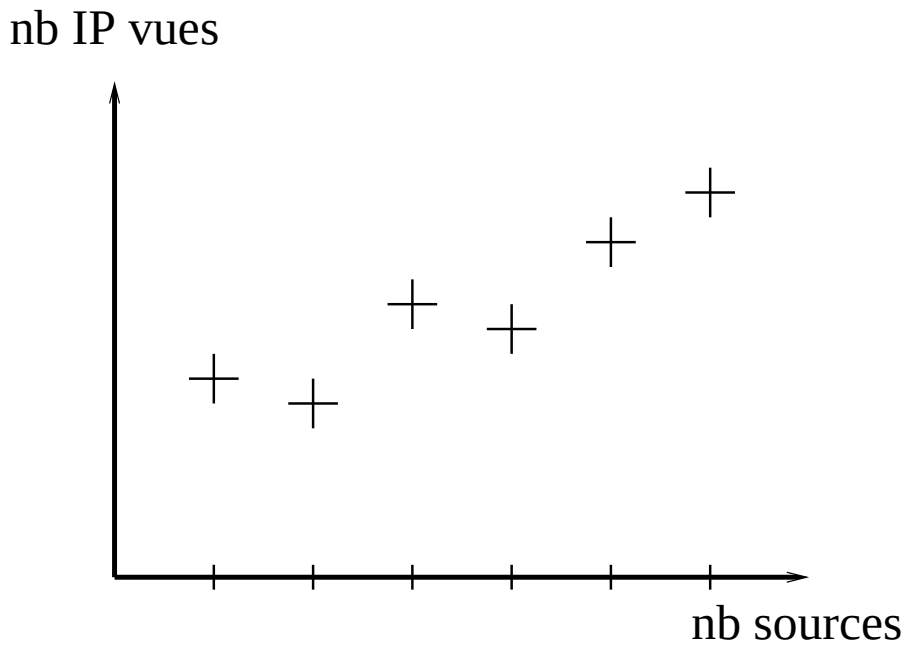
Méthodologie

Ce qu'on veut :



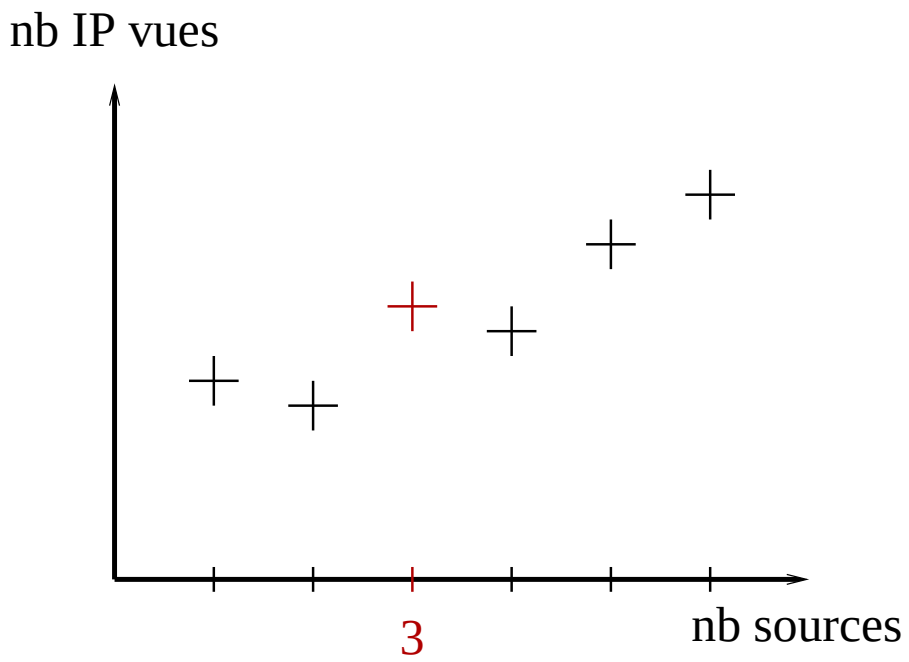
Méthodologie

Ce qu'on veut :



même chose pour les destinations

Problème



Nombre d'IP vues avec 3 sources : quelles 3 sources ?

Exemple

une source $\rightarrow$ ensemble des IP vues

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

$s_1 + s_3 + s_6 \rightarrow 5$ IP
$s_1 + s_4 + s_5 \rightarrow 10$ IP

Pas de choix naturel

Exemple

une source $\rightarrow$ ensemble des IP vues

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

$s_1 + s_3 + s_6 \rightarrow$5 IP
$s_1 + s_4 + s_5 \rightarrow$10 IP

Pas de choix naturel

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

1 sources : s_1

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

2 sources : $s_1 s_5$

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

3 sources : $s_1 s_5 s_4$

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

4 sources : $s_1 s_5 s_4 s_2$

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

5 sources : $s_1 s_5 s_4 s_2 s_3$

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

6 sources : $s_1 s_5 s_4 s_2 s_3 s_6$

Stratégie gloutonne

À chaque étape :

ajouter la source qui ajoute le plus d'informations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

sources : $s_1 s_5 s_4 s_2 s_3 s_6$

Motivation : "meilleur" des cas

Complexité

Union de deux ensembles

complexité minimum : taille du plus petit
(dépend de l'implémentation)

Deuxième étape

calcul de $n-1$ unions
$\rightarrow (n-1) \times k$ si tous les ensembles ont taille k.

À l'étape i

$n-i$ unions
$\rightarrow (n-i) \times k$

Complexité

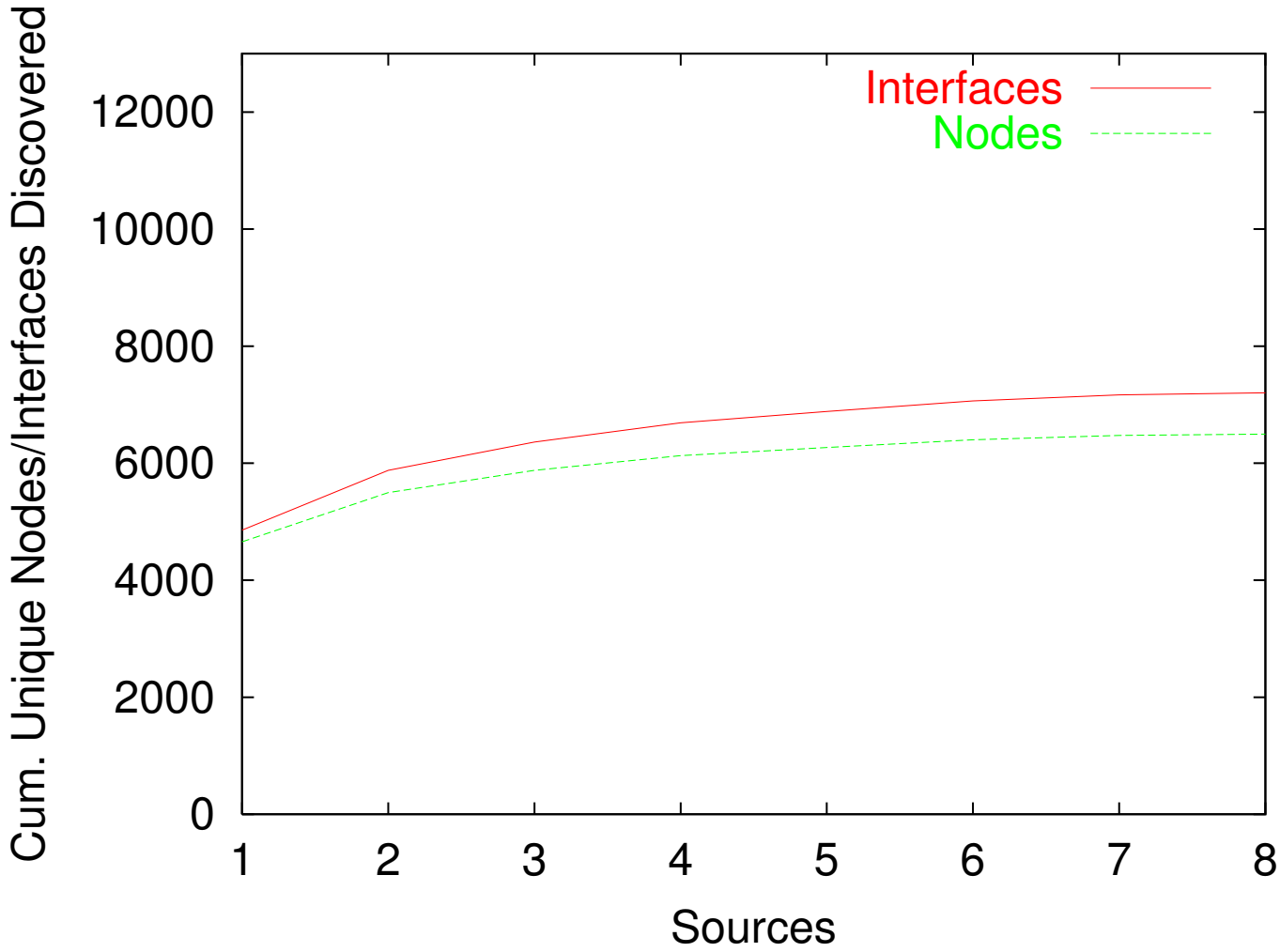
À l'étape i

$n - i$ unions
$\rightarrow (n - i) \times k$

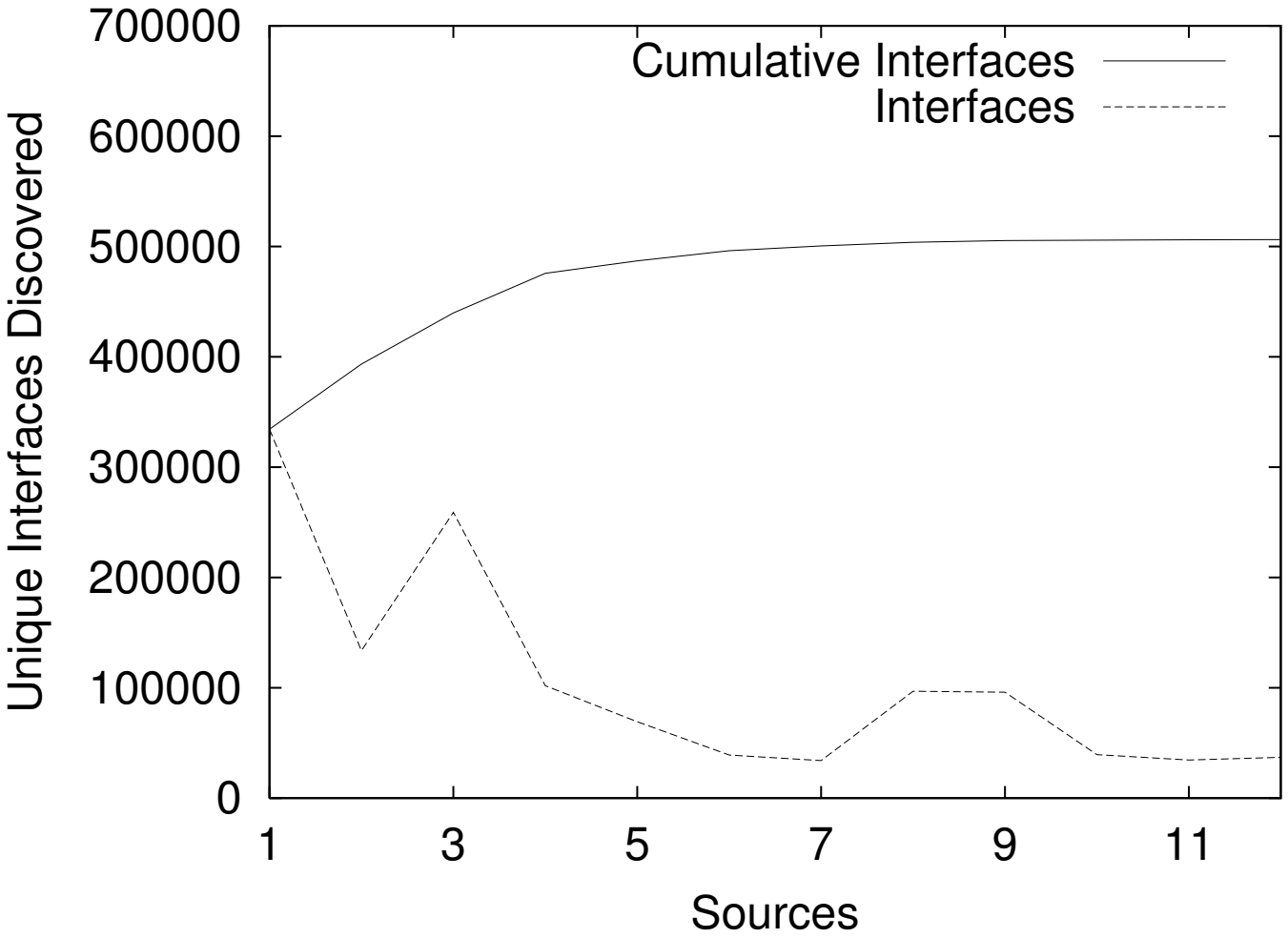
$$k(n - 1 + n - 2 + \ldots + 2 + 1)$$
$$= \frac{kn(n-1)}{2}$$
$$\mathcal{O}(kn^2)$$

Long si on a un grand nombre de sources

Résultats


Cum. Unique Nodes/Interfaces Discovered
12000
10000
8000
6000
4000
2000
0
1
2
3
4
5
6
7
8
Sources
Interfaces
Nodes

Résultats


Cumulative Interfaces
Interfaces
Unique Interfaces Discovered
700000
600000
500000
400000
300000
200000
100000
0
1
3
5
7
9
11
Sources

Observations

Convergence de la courbe :
les dernières sources n'apportent quasiment pas d'informations → marginal utility

à discuter plus tard

Utilité des destinations

Dans l'idéal, approche inverse :
Chaque destination $\rightarrow$ ensemble des IP vues

Stratégie gloutonne coûteuse
$\rightarrow$ stratégie aléatoire

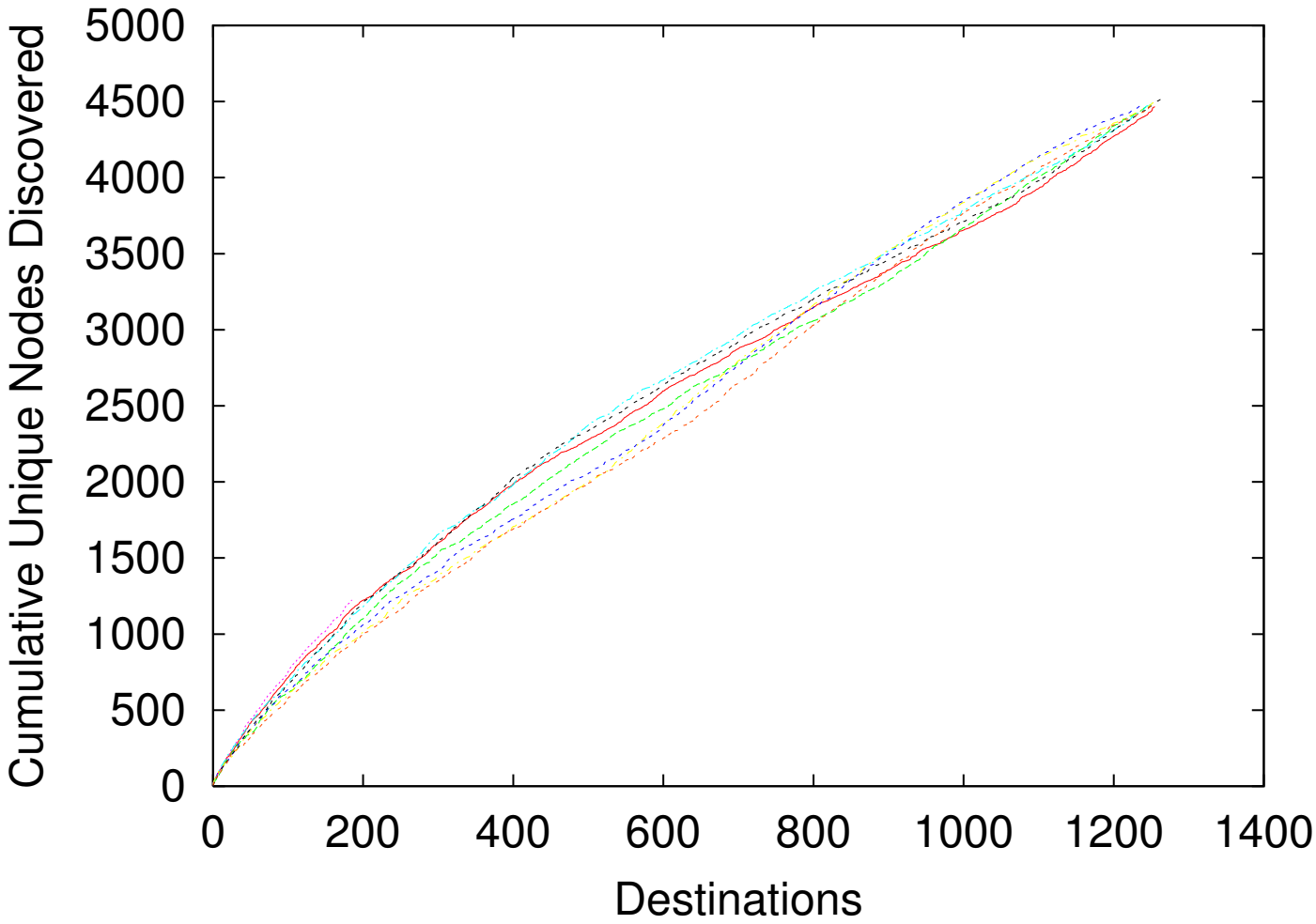
Pour une source

À chaque étape :

- Rajouter une destination au hasard

Comparer les courbes pour toutes les sources

Résultats


Cumulative Unique Nodes Discovered
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
1400
Destinations

Observations

Croissance linéaire :
apport *similaire* pour toutes les destinations

Différence entre sources et destinations

Différence entre les courbes
$\rightarrow$ différence entre les sources et les destinations ?

s sources, d destinations $\iff$ d sources, s destinations

$\rightarrow$ Importance de la stratégie utilisée pour les courbes
gloutonne, aléatoire

Différence entre sources et destinations

Différence entre les courbes
$\rightarrow$ différence entre les sources et les destinations ?

s sources, d destinations $\iff$ d sources, s destinations

$\rightarrow$ Importance de la stratégie utilisée pour les courbes
gloutonne, aléatoire

Différence entre sources et destinations

Différence entre les courbes
$\rightarrow$ différence entre les sources et les destinations ?

s sources, d destinations $\iff$ d sources, s destinations

$\rightarrow$ Importance de la stratégie utilisée pour les courbes
gloutonne, aléatoire

Critiques

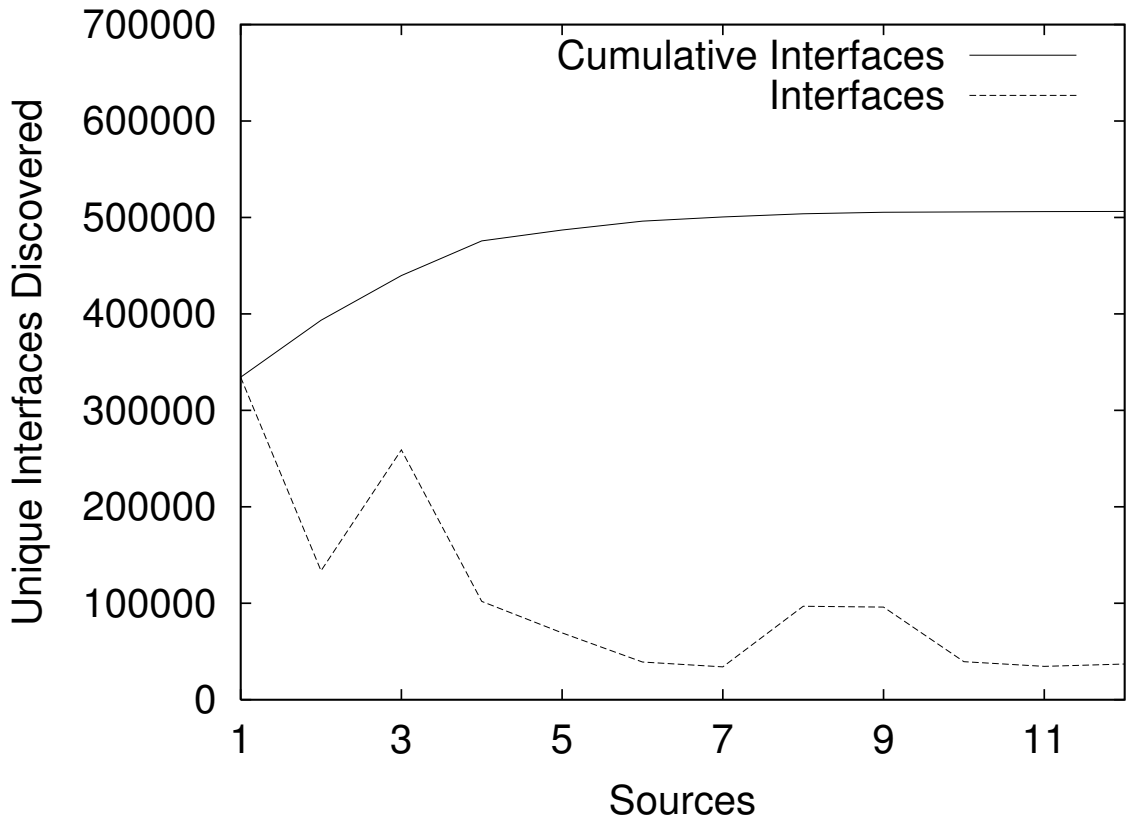
Papier intéressant, mais :

Manque de détails sur :

→ disparité entre les sources
(une source voit seulement 184 nœuds ($>$ 4000 pour la plus grande))
→ influence de la stratégie

Question : choix des sources plus important que le nombre ?

Critiques



Dernières sources : peu d'apport en soi

La stratégie gloutonne conditionne l'allure de la courbe

pas de stratégie naturelle

Pour mieux comprendre

Comparer les différentes stratégies

Données

[Ouédraogo, Magnien, 2009]

Données

- 11 sources
- 3 000 destinations
- 100 traceroutes par jour
- $\sim$ 2 mois

Différence entre les sources

Nombre d'IP vues par sources

Varie entre :

- $\sim 16\,500$
- $\sim 26\,500$

Toutes les sources ne sont pas équivalentes

Influence des sources ou des destinations

Trois stratégies naturelles

- gloutonne
- aléatoire
- gloutonne-min

  ajouter la source qui apporte le moins d'information

Influence des sources ou des destinations

Stratégie gloutonne $\neq$ maximum possible avec k sources

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, e, f\}$
$s_3 : \{a, c, d, g\}$

Influence des sources ou des destinations

Stratégie gloutonne $\neq$ maximum possible avec k sources

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, e, f\}$
$s_3 : \{a, c, d, g\}$

1 sources : s_1

Influence des sources ou des destinations

Stratégie gloutonne $\neq$ maximum possible avec k sources

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, e, f\}$
$s_3 : \{a, c, d, g\}$

2 sources : $s_1 s_2$

Influence des sources ou des destinations

Stratégie gloutonne $\neq$ maximum possible avec k sources

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, e, f\}$
$s_3 : \{a, c, d, g\}$

3 sources : $s_1 s_2 s_3$

Influence des sources ou des destinations

Stratégie gloutonne $\neq$ maximum possible avec k sources

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, e, f\}$
$s_3 : \{a, c, d, g\}$

3 sources : $s_1 s_2 s_3$

$$s_2 + s_3 : 7 \text{ IP}$$

Représentativité du maximum ?
Coût du calcul du maximum

Influence des sources ou des destinations

Autres stratégies

- Max → max sur 1000 ordres aléatoires
- Min → min sur 1000 ordres aléatoires
- Aléatoire → moyenne sur 1000 ordres aléatoires

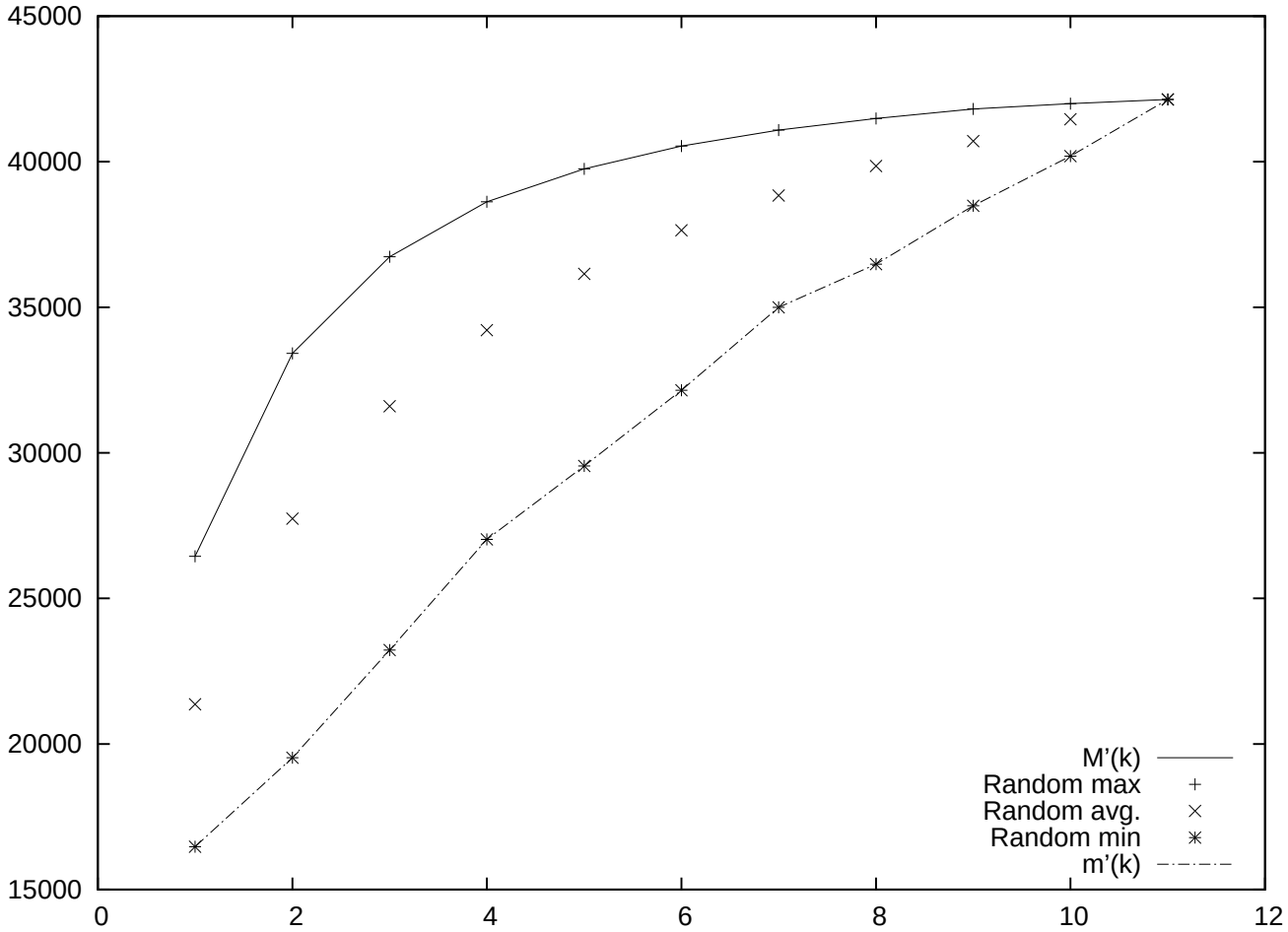
Influence des sources ou des destinations

Exemple

$s_1 : \{a, b, c, d, e\}$
$s_2 : \{a, b, c, d, f\}$
$s_3 : \{a, b\}$
$s_4 : \{g, h\}$
$s_5 : \{i, j, k\}$
$s_6 : \{a, d\}$

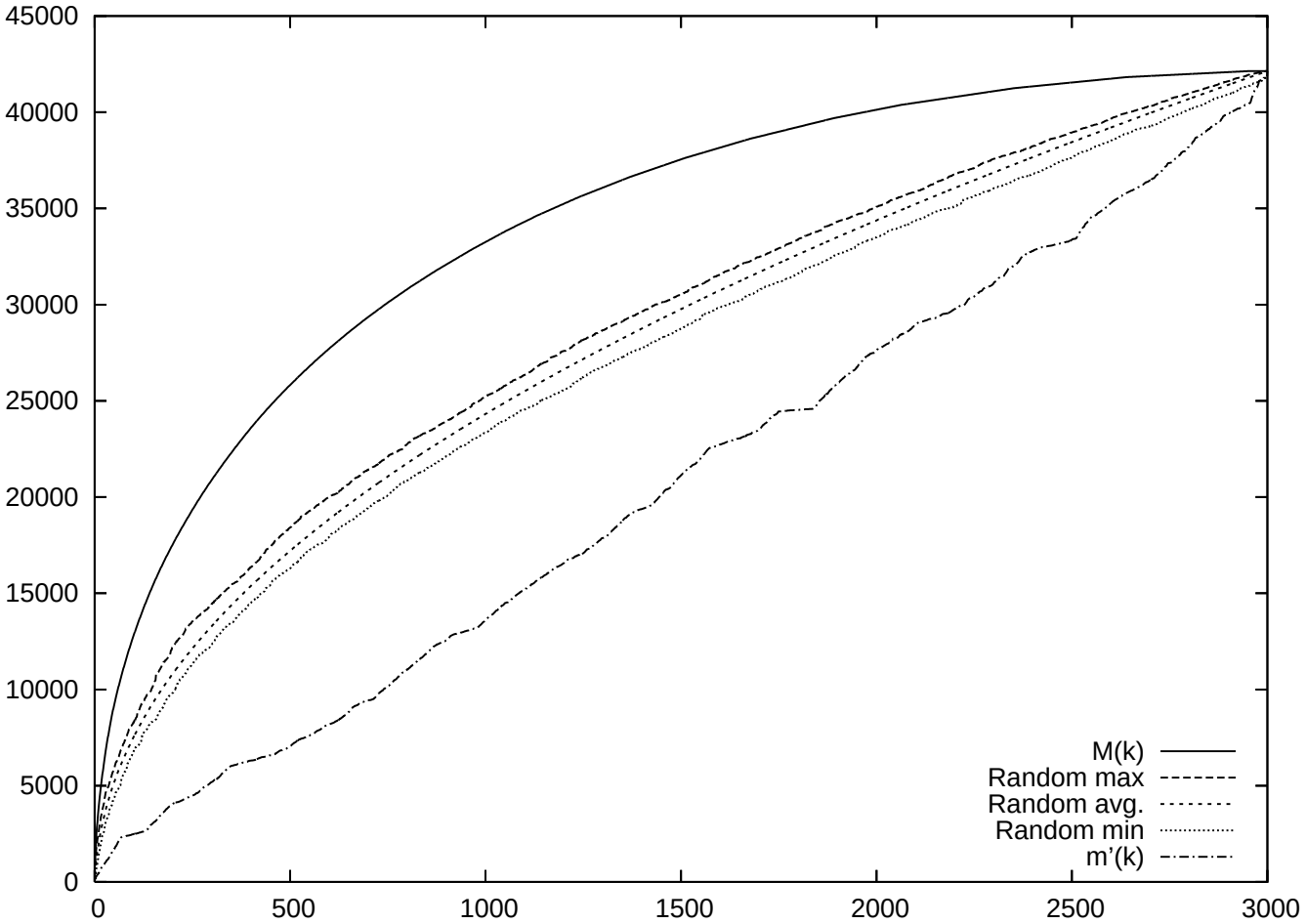
	s_3	s_4	s_6	s_5	s_2	s_1
	2	4	5	8	10	11
	s_5	s_6	s_2	s_4	s_3	s_1
	3	5	7	9	10	11
Min	2	4	5	8	10	11
Max	3	5	7	9	10	11
Moyenne	2.5	4.5	6	8.5	10	11

Résultats



Influence des sources

Résultats



Influence des destinations

Observations

- Toutes les courbes finissent au même point : n
- Glouton + moyenne : comportements similaires pour les sources et les destinatiosn
- Variabilité plus grande en pratique pour les sources

Peu de sources

Conclusion

Utilité diminue, ne devient pas nulle

Choix des sources peut-être plus important que le nombre

Outline

Biais lié à l'exploration

[Sampling Biases in IP Topology Measurements
Lakhina, Byers, Crovella, Xie, 2003]

Principe : simulations

- Générer des graphes $\rightarrow$ topologie
- Simuler des traceroute $\rightarrow$ mesure
- Observer les résultats

Caractère explicatif

Implémentation

Générer des graphes

- Aléatoires $\rightarrow$ Erdös-Rényi
- Degrés fixés $\rightarrow$ modèle de configurations

Implémentation – traceroute

Comment simuler traceroute ?
Plusieurs possibilités

Implémentation – traceroute

Comment simuler traceroute ?
Plusieurs possibilités

Choix courants

- route = plus court chemin (faux mais on n'a pas mieux)

Plus court chemin

- Un seul/tous les plus courts chemins ?
- Si un seul, lequel ?

Choix des auteurs

Associer un poids à chaque lien ($\rightarrow$ graphe pondéré)

$$1 + \epsilon, \epsilon \in [-1/n, 1/n]$$

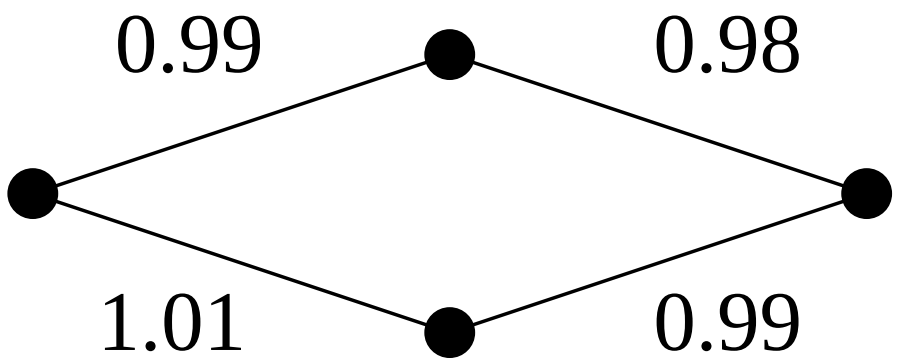
Longueur d'un chemin : somme des poids des liens
$\rightarrow$Tous les chemins ont des longueurs différentes

Choix des auteurs

Associer un poids à chaque lien ($\rightarrow$ graphe pondéré)

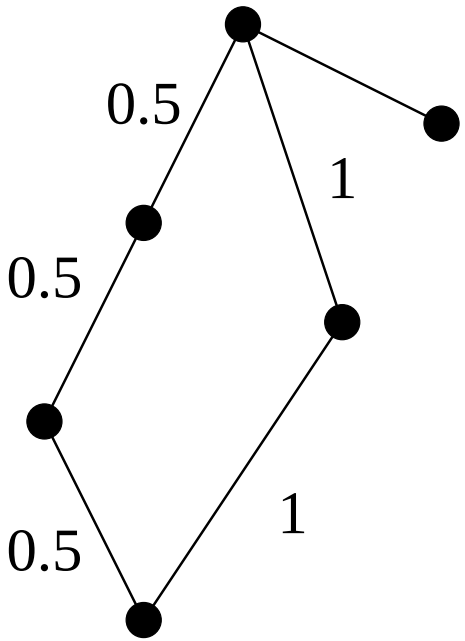
$$1 + \epsilon, \epsilon \in [-1/n, 1/n]$$

Longueur d'un chemin : somme des poids des liens
$\rightarrow$Tous les chemins ont des longueurs différentes



Calcul du plus court chemin pondéré

Parcours en largeur pas adapté



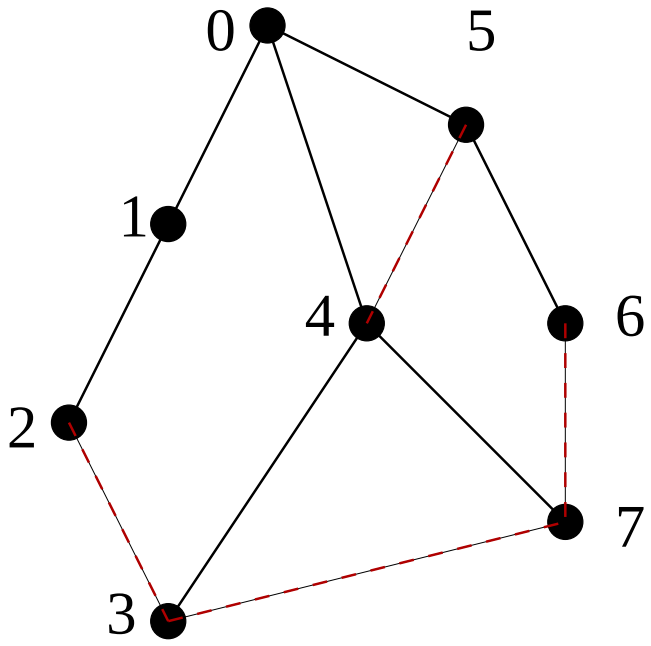
→ Algorithme de Dijkstra (pas détaillé ici)

Notre choix – parcours en largeur

Pas de poids
Distances calculées par parcours en largeur
Stockage de la sortie d'un parcours → tableau

La case i contient le père de i
Le père de la racine est la racine elle-même

0	0	1	4	0	0	5	4
0	1	2	3	4	5	6	7



Restreindre aux destinations

Tableau initialisé à -1

Pour chaque destination `d` :

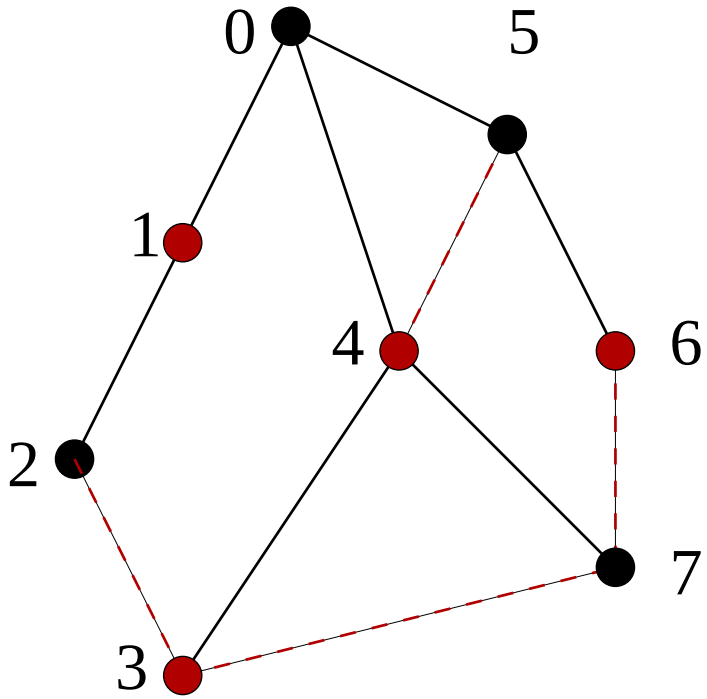
- Tant que `AR[d] == -1`
  - AR[d] = A[d]
  - d = A[d]

A

0	0	1	4	0	0	5	4
0	1	2	3	4	5	6	7

AR

0	0	-1	4	0	0	5	-1
0	1	2	3	4	5	6	7



Calcul des degrés

Degré d'un nœud dans l'arbre des parcours en largeur :
nombre de fois où il apparaît + 1

(racine : nombre de fois -1)

0	0	-1	4	0	0	5	-1
0	1	2	3	4	5	6	7

(cases à -1 : nœuds qui ne sont pas dans l'arbre)

Plusieurs sources

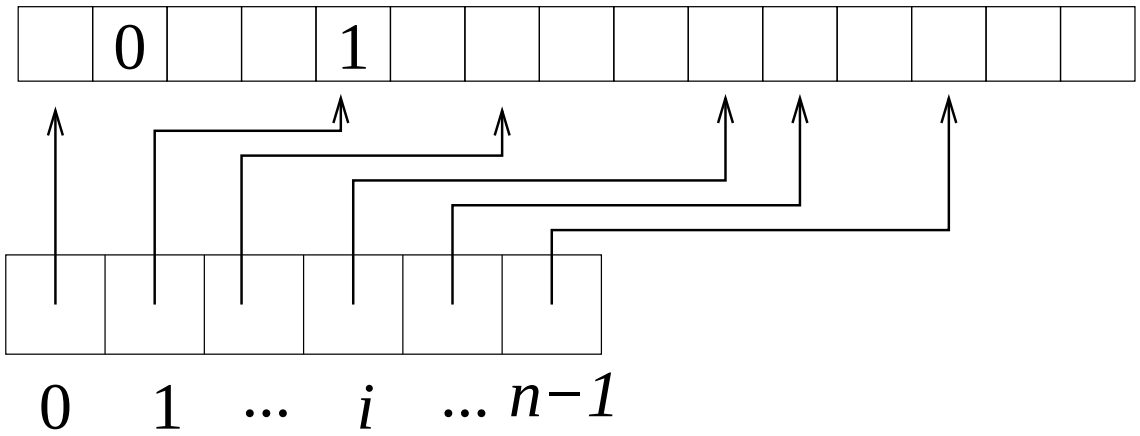
Plusieurs sources :
un parcours en largeur par source

Comment calculer le degré des nœuds ?

Marquer les liens comme présents ou absents

Plusieurs sources

Allocation d'une structure graphe identique



- 0 : lien absent
- 1 : lien présent

Pour chaque lien dans chaque parcours en largeur :
$\rightarrow$ passer le lien à 1 dans la copie

Connexité

Problème si le graphe n'est pas connexe

Plusieurs solutions

- Choisir les sources et destinations dans la même composante connexe
- Garder uniquement les chemins qui existent
- Utiliser uniquement des graphes connexes

Pas de solution parfaite

Connexité

Problème si le graphe n'est pas connexe

Choix des auteurs :

se restreindre à la plus grande composante connexe

Simulations

Examen de deux hypothèses

Graphes d'Erdös-Rényi (degrés homogènes)

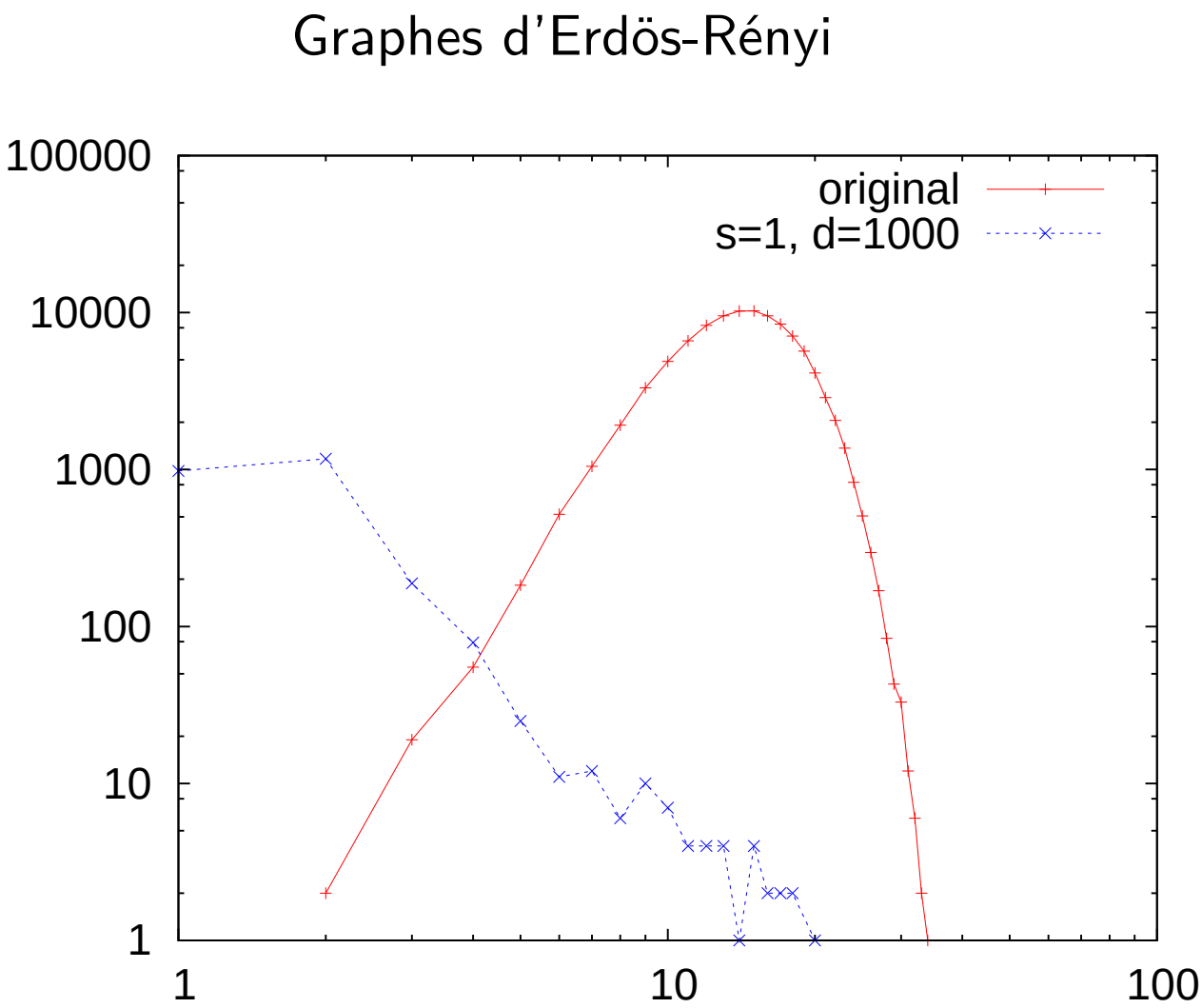
- $n = 100\,000$
- $m = 750\,000$ ($d^\circ(G) = 15$)
- sources : 1, 5, 10
- destinations : 1000

Choisies au hasard

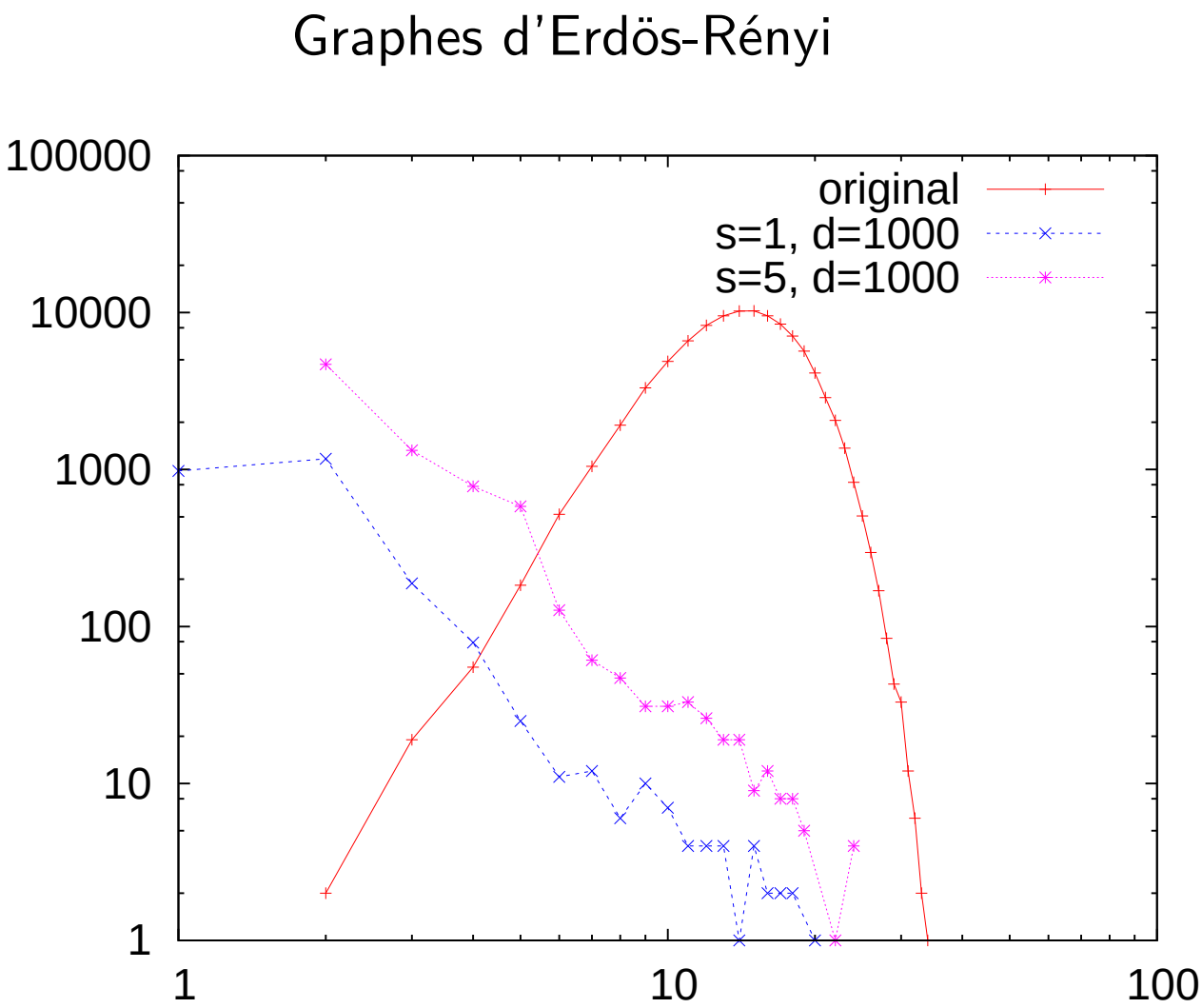
Distribution des degrés fixés (hétérogène)

- $n \sim 100\,000$
- $m \sim 190\,000$
- loi de puissance, $\alpha \sim 2.1$

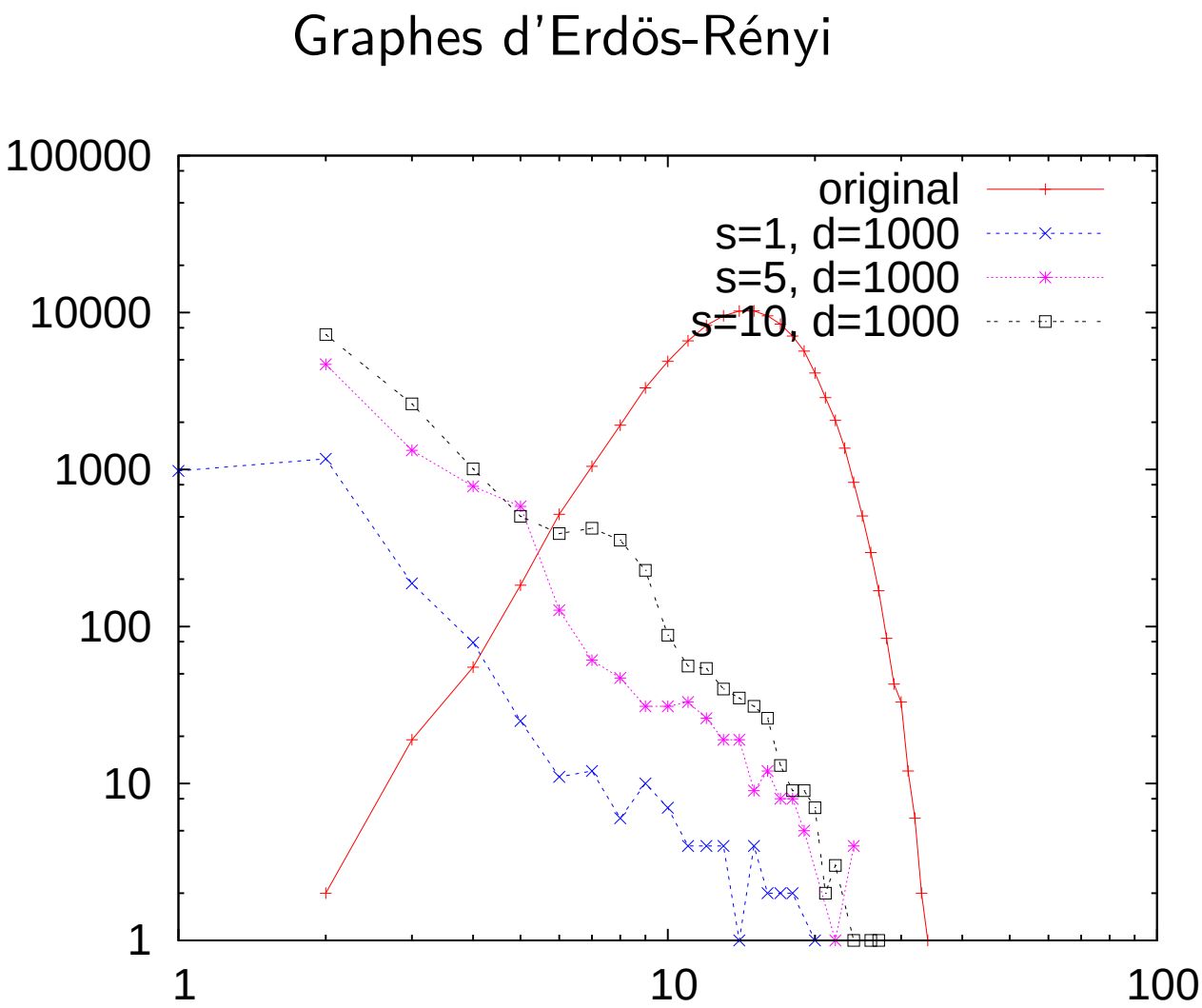
Résultats


Graphes d'Erdös-Rényi
original
s=1, d=1000
100000
10000
1000
100
10
1
1
10
100

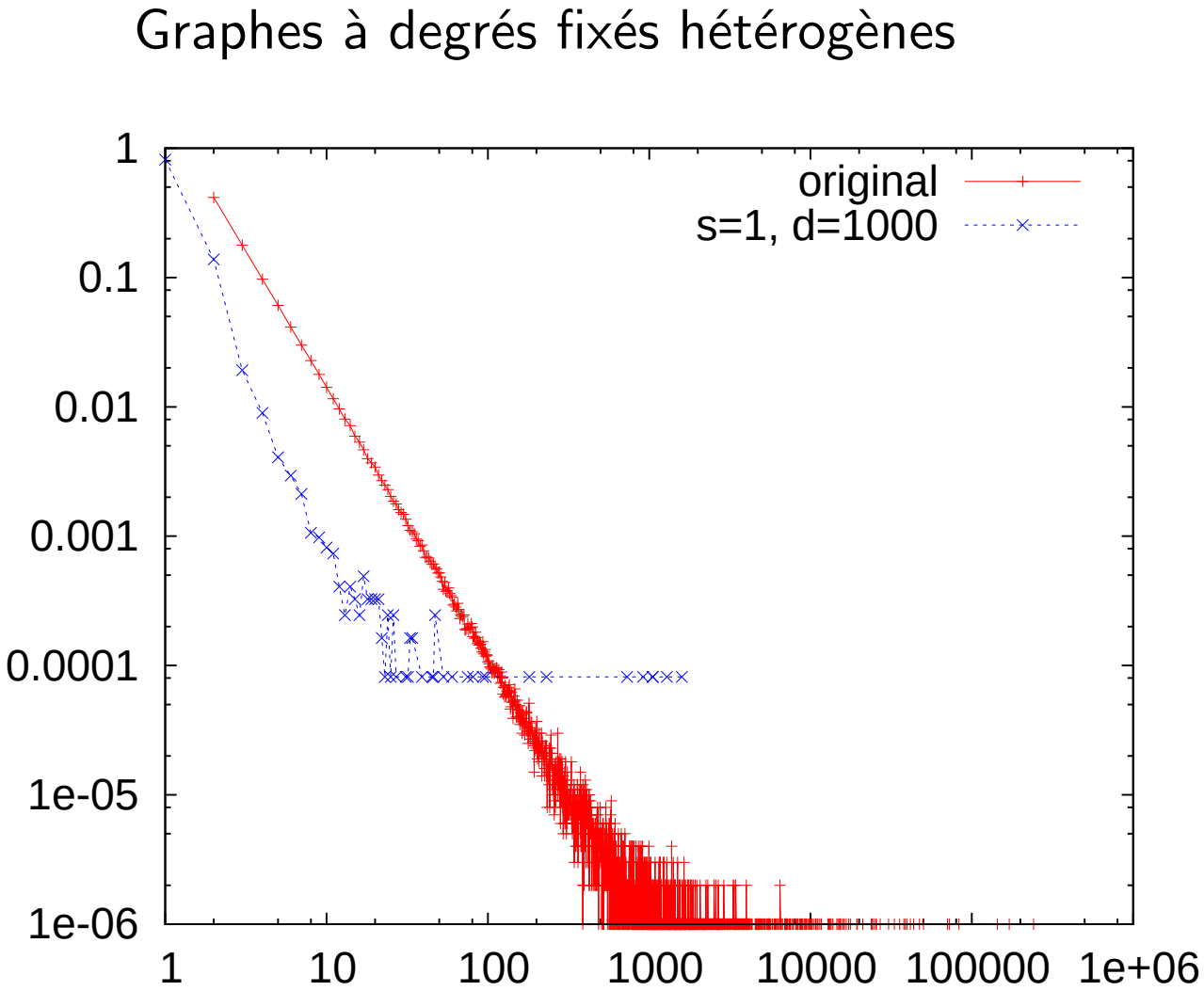
Résultats


Graphes d'Erdös-Rényi
original
s=1, d=1000
s=5, d=1000
100000
10000
1000
100
10
1
1
10
100

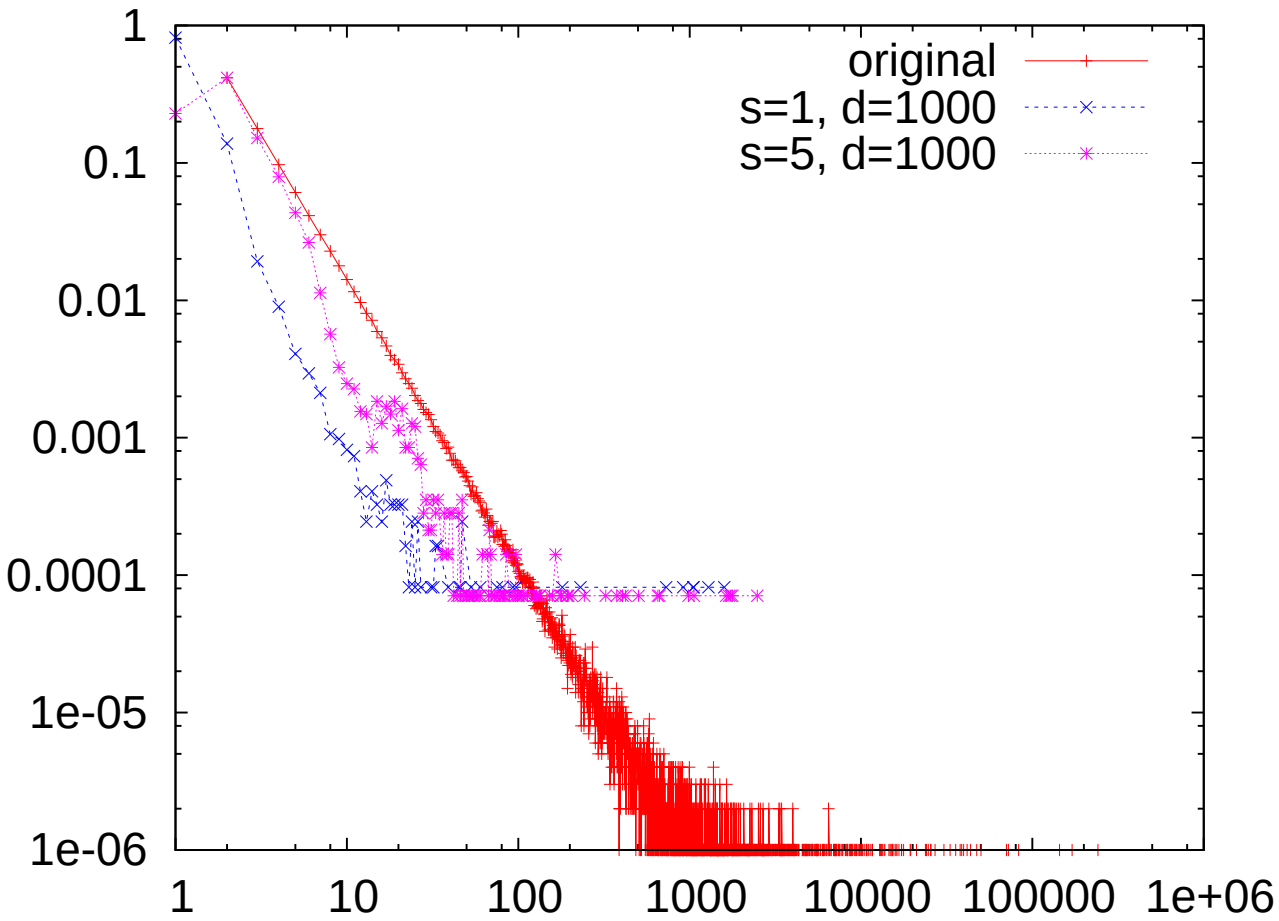
Résultats


Graphes d'Erdös-Rényi
100000
10000
1000
100
10
1
1
10
100
original
s=1, d=1000
s=5, d=1000
s=10, d=1000

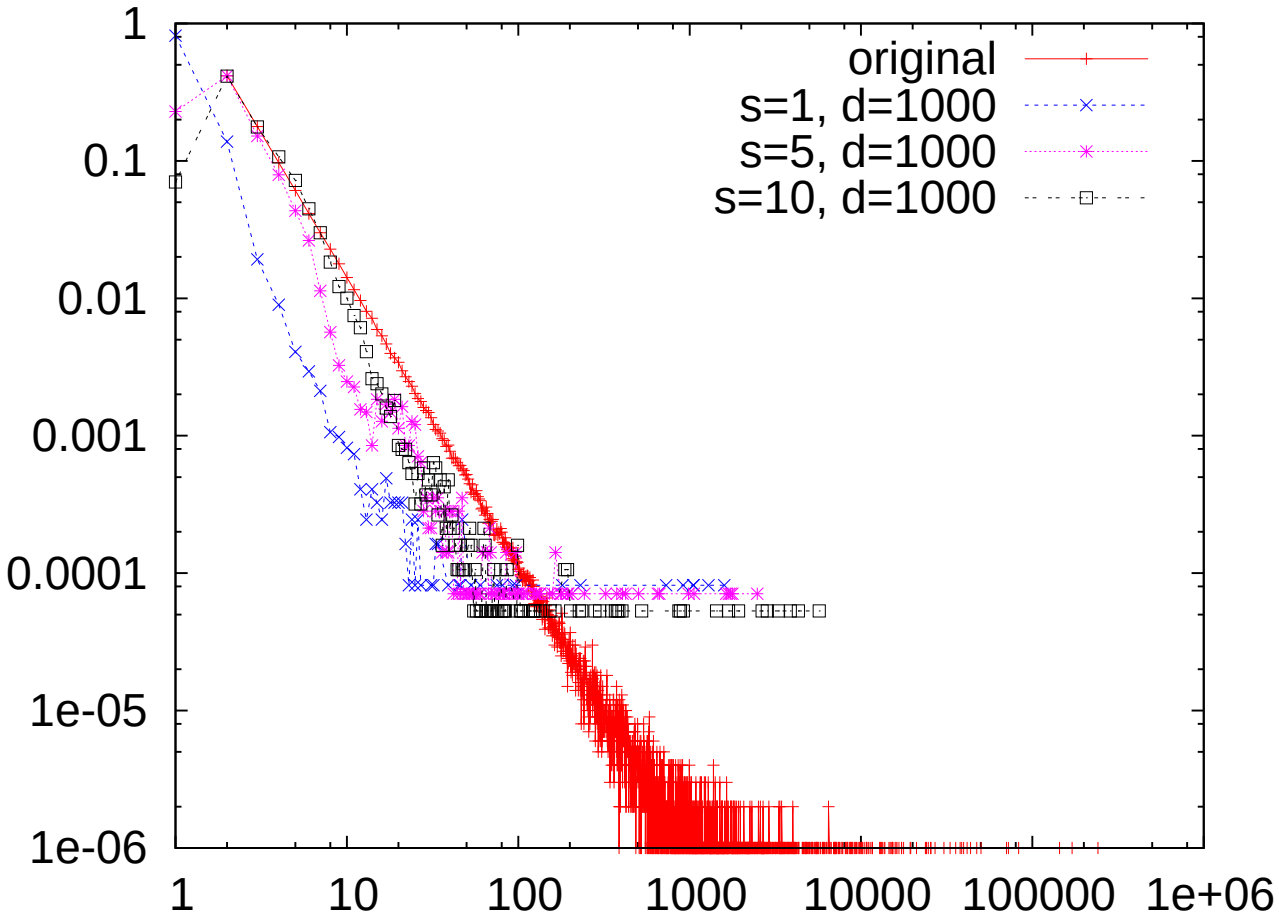
Résultats


Graphes à degrés fixés hétérogènes
original
s=1, d=1000
1
0.1
0.01
0.001
0.0001
1e-05
1e-06
1
10
100
1000
10000
100000
1e+06

Résultats


Graphes à degrés fixés hétérogènes
original
s=1, d=1000
s=5, d=1000
1
0.1
0.01
0.001
0.0001
1e-05
1e-06
1
10
100
1000
10000
100000
1e+06

Résultats


Graphes à degrés fixés hétérogènes
original
s=1, d=1000
s=5, d=1000
s=10, d=1000
1
0.1
0.01
0.001
0.0001
1e-05
1e-06
1
10
100
1000
10000
100000
1e+06

Observations

- Distribution observée $\neq$ distribution réelle
- Graphes aléatoires : différence qualitative

  homogène $\rightarrow$ hétérogène
- Graphes à degrés fixés : différence quantitative

  pente, degré max, ...

Attention :
Graphes aléatoires : Degré maximum observé ~ 30
$\rightarrow$impossible de conclure sur l'hétérogénéité

Observations

- Distribution observée $\neq$ distribution réelle
- Graphes aléatoires : différence qualitative
  homogène $\rightarrow$ hétérogène
- Graphes à degrés fixés : différence quantitative
  pente, degré max, ...

Attention :
Graphes aléatoires : Degré maximum observé ~ 30
$\rightarrow$impossible de conclure sur l'hétérogénéité

Conclusion de l'article

On peut avoir une difference
Distribution observée hétérogène $\not\Longleftrightarrow$ distribution réelle hétérogène

Pas de conclusion sur la distribution réelle

Discussion (1/2)

Résultat très important

- D'un point de vue théorique
- Besoin de faire attention en pratique

Quelles conclusions tirer en pratique ?

Distribution observée hétérogène

$\rightarrow$ Distribution réelle homogène ?
$\rightarrow$ Distribution réelle hétérogène ?

Discussion (2/2)

Cas de graphes aléatoires :
Degré maximal observé :
proche du degré moyen du graphe.

En pratique, degré maximum observé > 1000
$\rightarrow$ graphe aléatoire de degré moyen $= 1000$?

$\rightarrow$distribution réelle probablement hétérogène
Besoin de plus d'études

Sources du biais

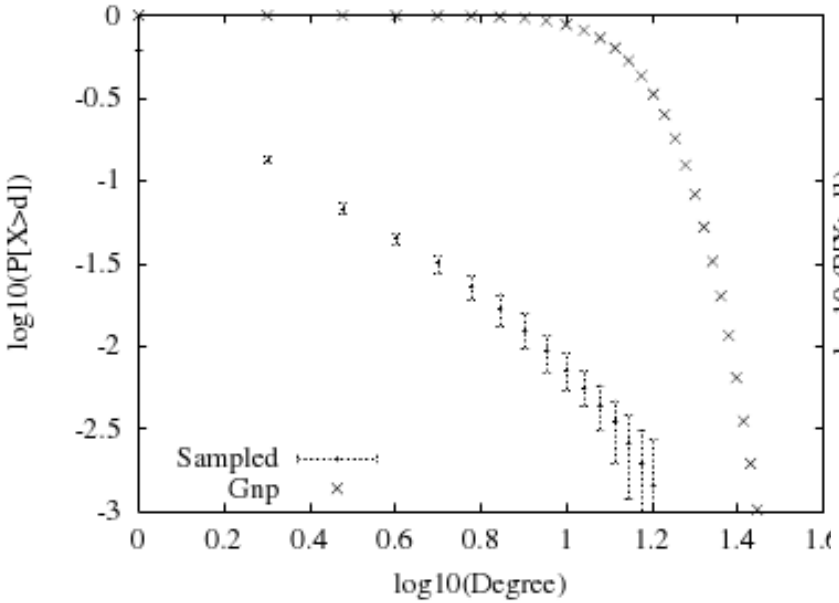
Biais dans l'échantillon des noeuds?

Pour chaque nœud : comparer le degré observé au vrai degré

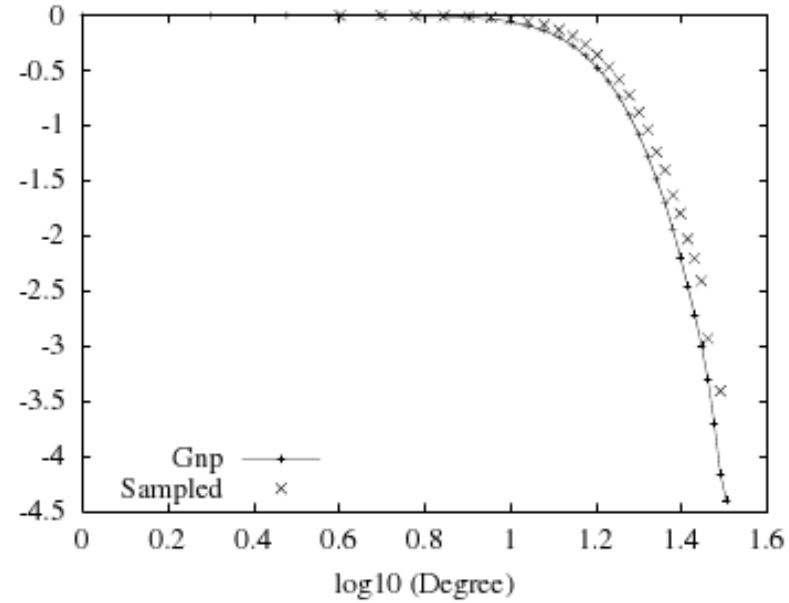
Sources du biais

Biais dans l'échantillon des noeuds?

distrib. observée / distrib. du graphe



distrib. des vrais degrés / distrib. du graphe

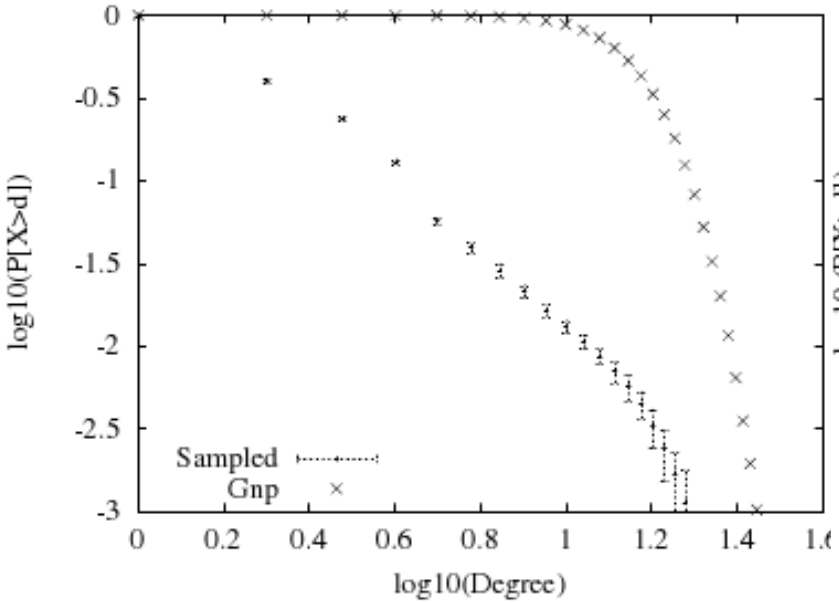


Avec 1 source

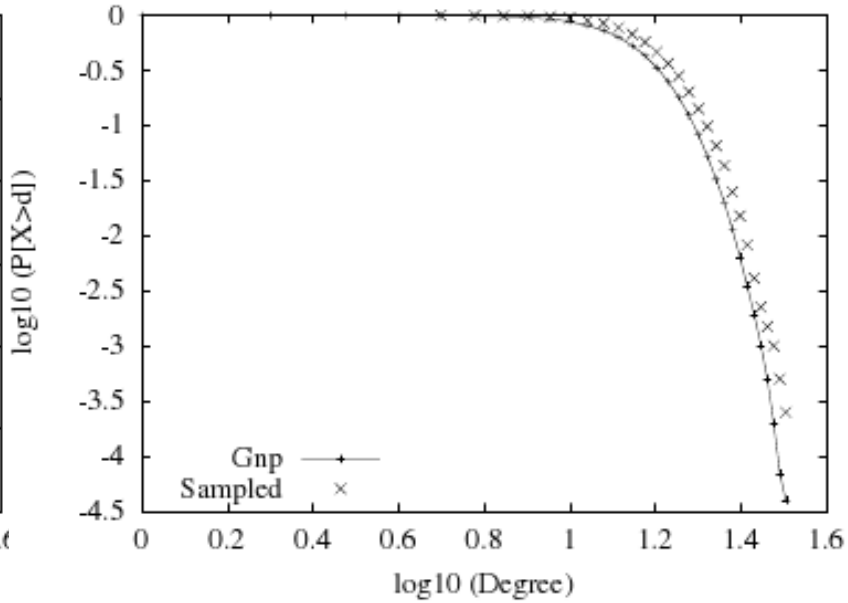
Sources du biais

Biais dans l'échantillon des noeuds?

distrib. observée / distrib. du graphe



distrib. des vrais degrés / distrib. du graphe

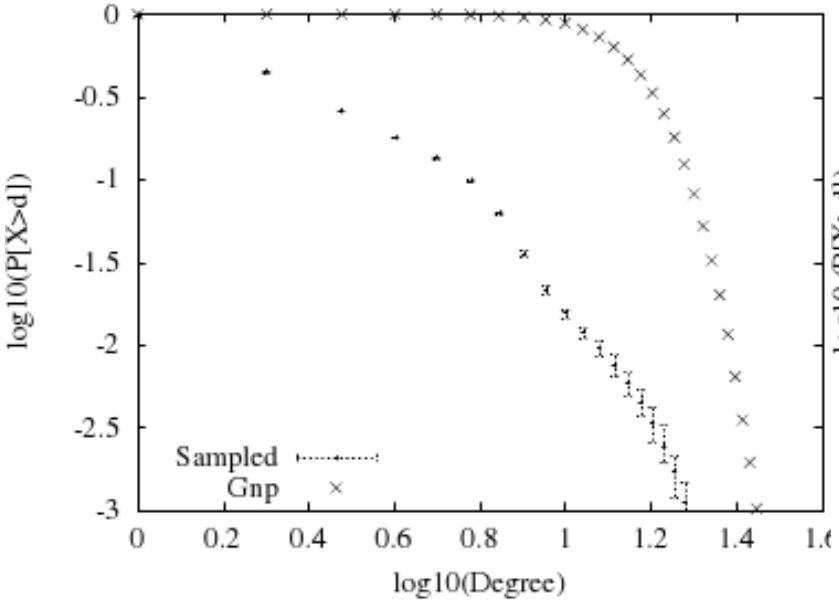


Avec 5 sources

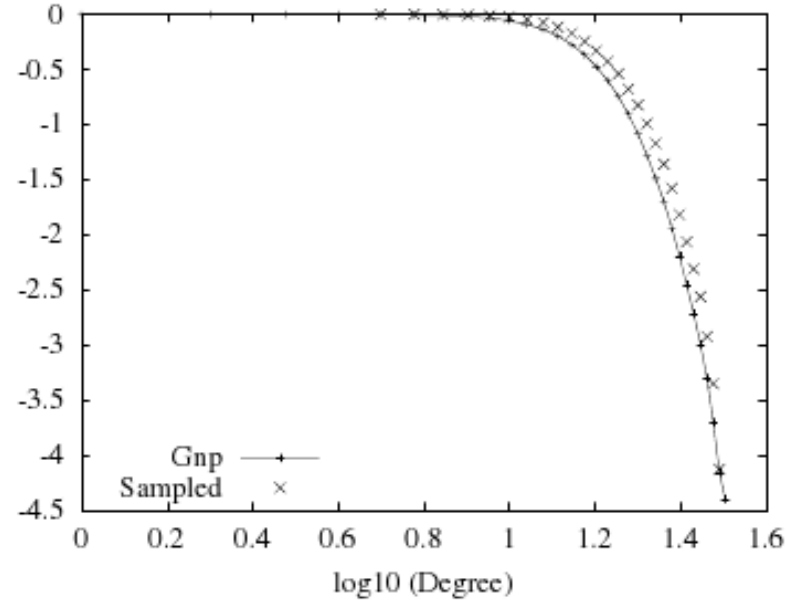
Sources du biais

Biais dans l'échantillon des noeuds?

distrib. observée / distrib. du graphe



distrib. des vrais degrés / distrib. du graphe



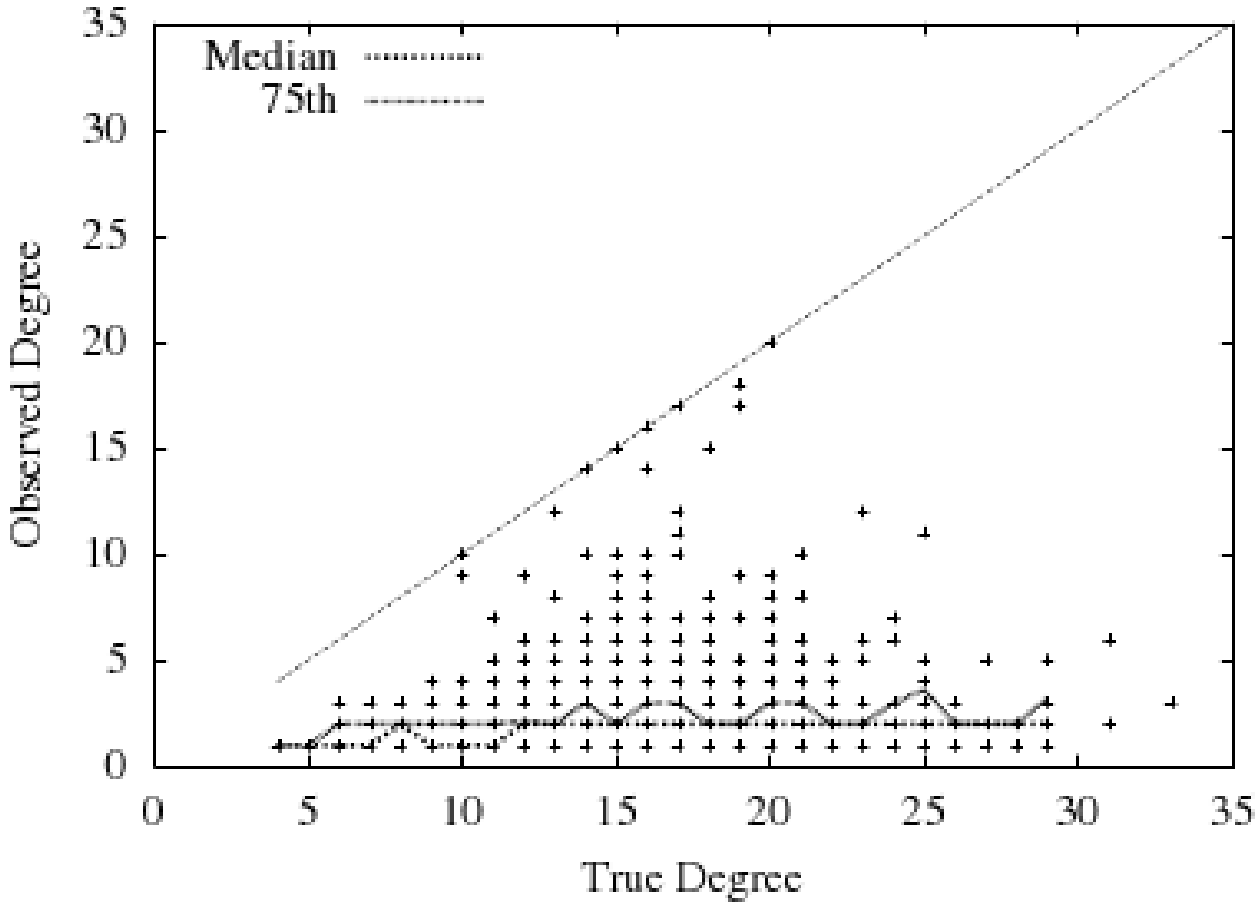
Avec 10 sources

Les nœuds sont choisis sans biais sur le degré

Sources du biais

Biais dans l'échantillon des liens?

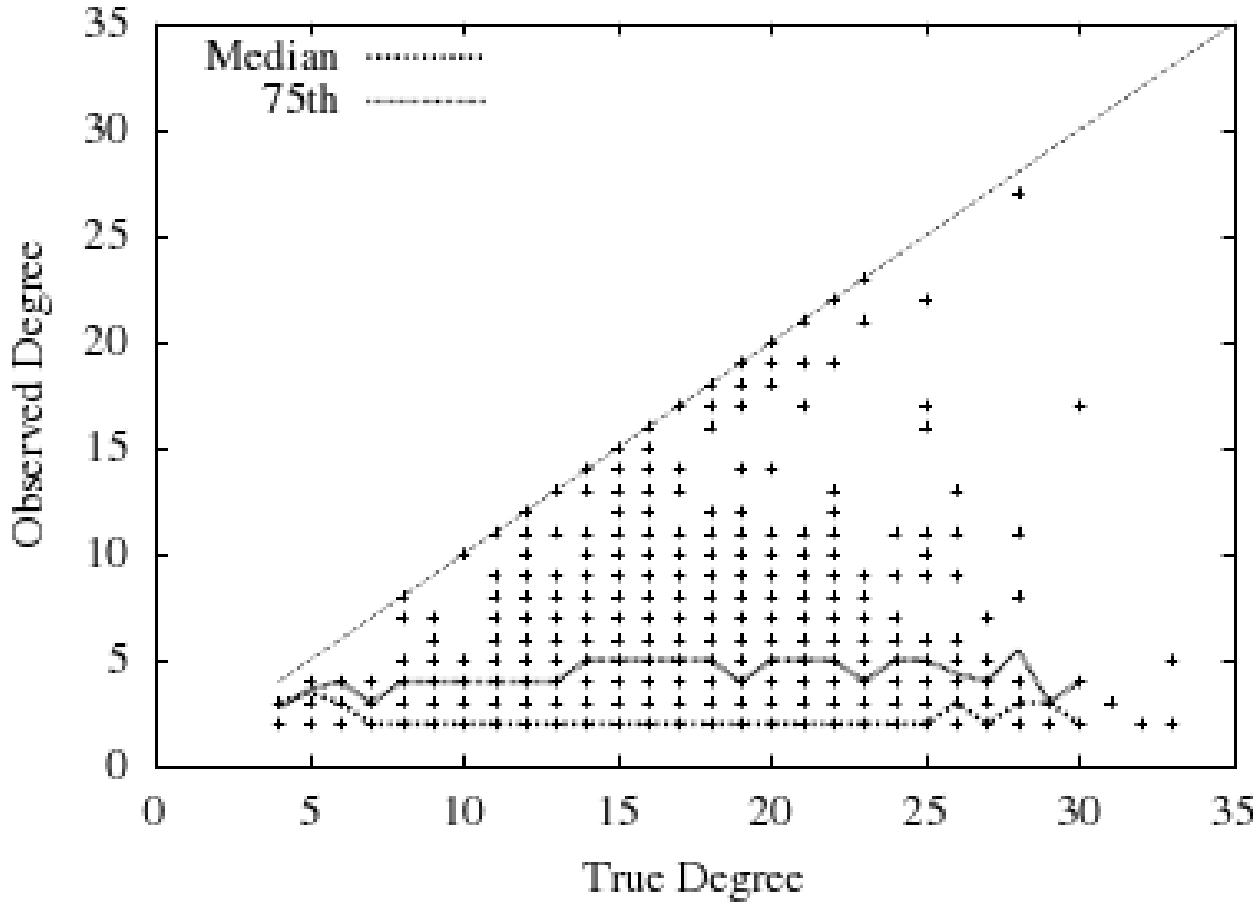
degré observé vs degré réel
Avec 1 source



Sources du biais

Biais dans l'échantillon des liens?

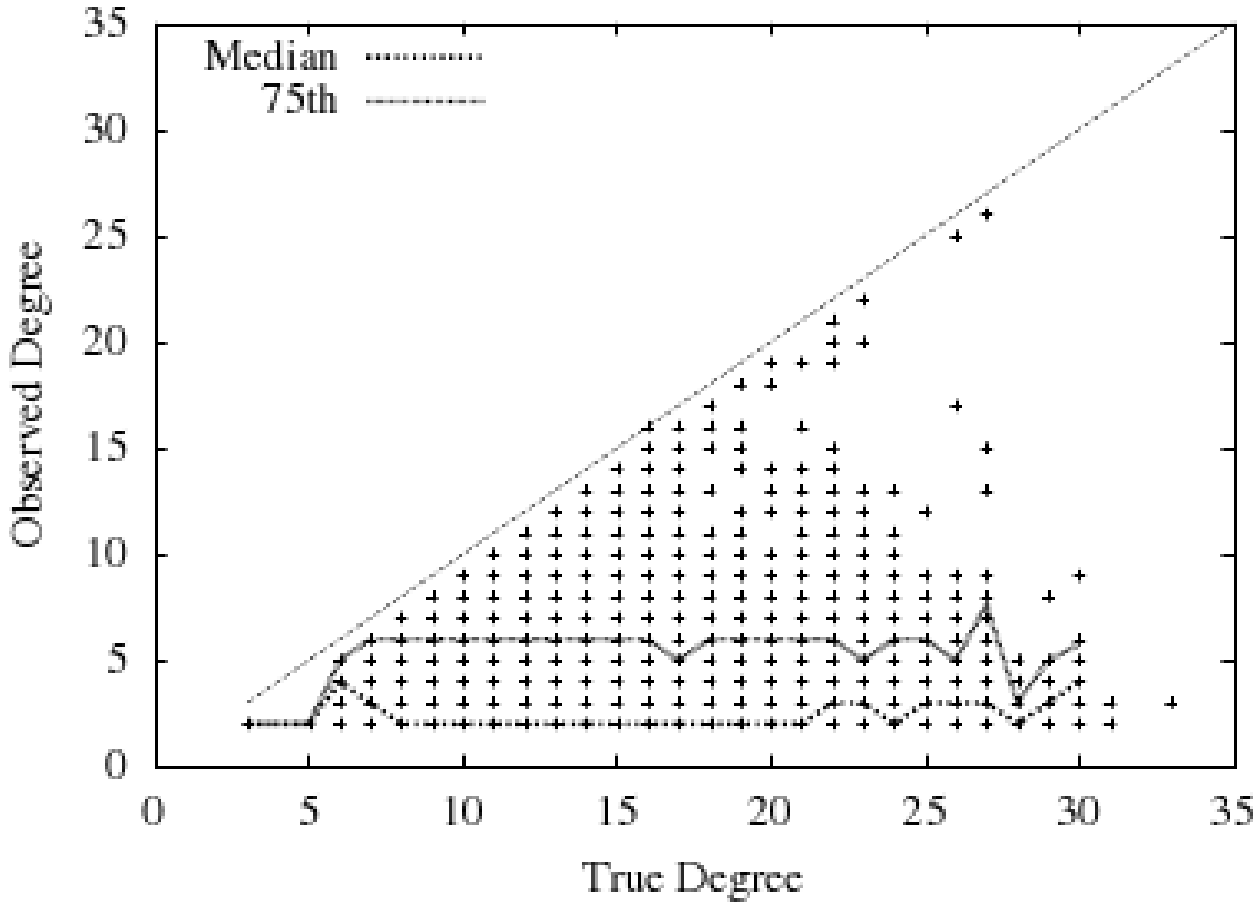
degré observé vs degré réel
Avec 5 sources



Sources du biais

Biais dans l'échantillon des liens?

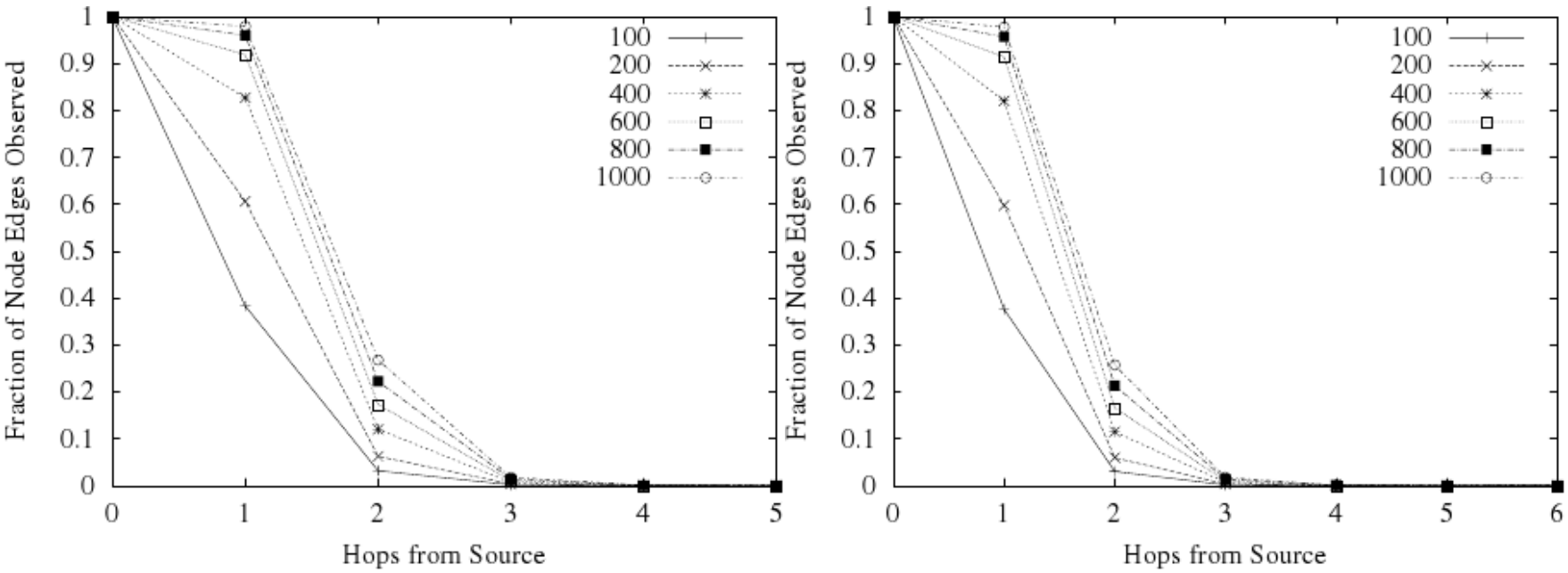
degré observé vs degré réel
Avec 10 sources



Sources du biais

Visibilité des arêtes en fonction de leur distance à la source

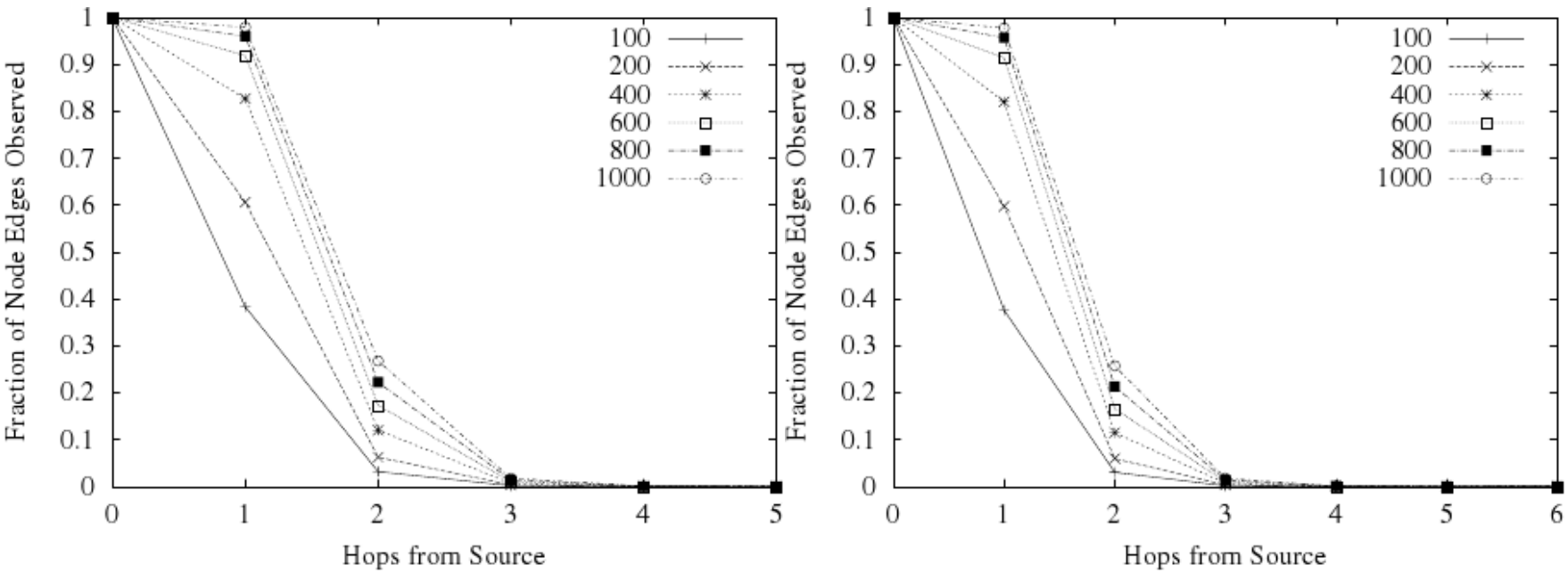
gauche : 10 000 sommets droite : 1 000 000 sommets



Sources du biais

Visibilité des arêtes en fonction de leur distance à la source

gauche : 10 000 sommets droite : 1 000 000 sommets



Plus une arête est loin de la source,
moins elle a de chances d'être vue

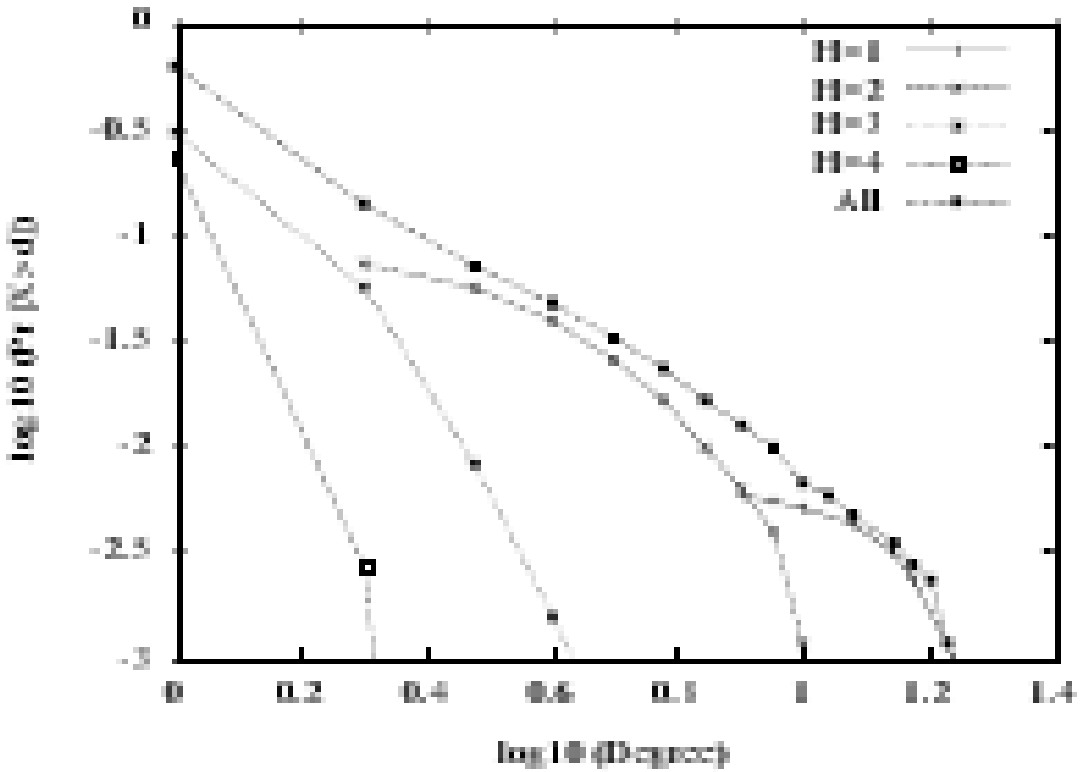
Échantillon donné $\longrightarrow$ biais ?

Étant donné un échantillon (et pas le graphe original), peut-on savoir s'il y a du biais ?

Échantillon donné → biais ?

Étant donné un échantillon (et pas le graphe original), peut-on savoir s'il y a du biais ?

Probabilité d'avoir degré d *et* distance h



Les nœuds les plus éloignés ont les plus faibles degrés