Graphes, réseaux et internet

Clémence Magnien

clemence.magnien@lip6.fr

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

avec Matthieu Latapy, Frédéric Ouédraogo, Guillaume Valadon, Assia Hamzaoui, ...

Outline

1. Graphes de terrain
   - Notions de graphes
   - Graphes de terrain
   - Problématiques
   - Dynamique
2. Dynamique - Mesure
3. Dynamique - Analyse
   - Analyse de graphes statiques
   - Nouvelles propriétés
   - Détection d'événements
   - Visualisation

Graphes de terrain

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

sciences sociales : collaboration, amitié, contacts sexuels, échanges, économie, etc

biologie : cerveau, gènes, protéines, écosystèmes, etc

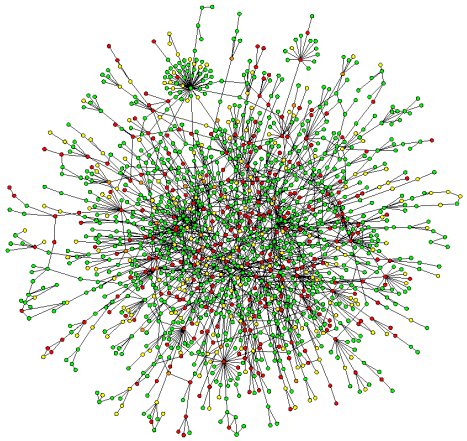
linguistique : synonymie, co-occurrence, etc

transport : routier, aérien, électrique, etc

etc, etc

réseaux de relations

contextes très différents



Graphes

graphe = *n* nœuds, *m* liens

capture la structure, la *topologie*

pas de *meilleur* dessin

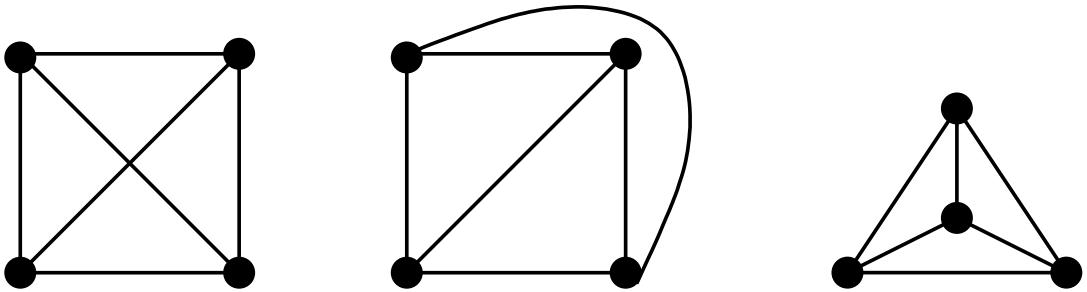
Graphes

graphe = *n* nœuds, *m* liens

capture la structure, la *topologie*

Graphe $\neq$ dessin : un graphe, plusieurs dessins

G = a, b, c, d ; a–b, a–c, a–d, b–c, b–d, c–d



pas de *meilleur* dessin

Propriétés de graphes

densité du graphe, δ

= probabilité d'existence de tout lien

= à quel point tout le monde est lié

$$\delta = \frac{2m}{n(n-1)}$$

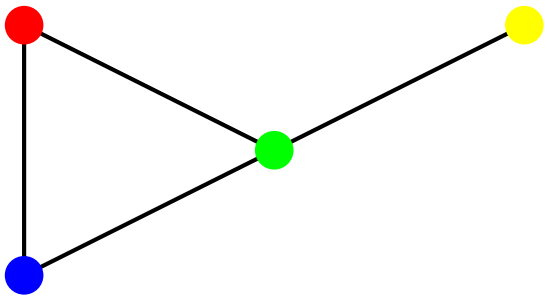
Propriétés de graphes

densité du graphe, δ

= probabilité d'existence de tout lien

= à quel point tout le monde est lié

$$\delta = \frac{2m}{n(n-1)}$$



$n = 4, m = 4, \delta = \frac{8}{12} = 0.66..$

Propriétés de graphes

degré d'un nœud v, $d^o(v)$

= nombre de liens de v

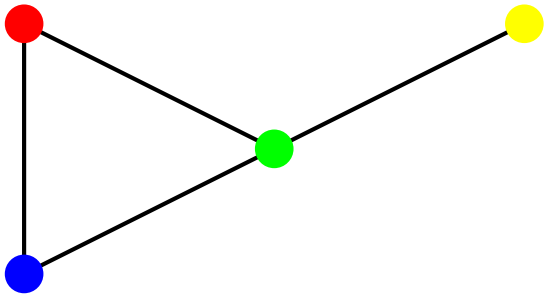
degré moyen $d^o = \frac{2m}{n}$

Propriétés de graphes

degré d'un nœud v, $d^o(v)$

= nombre de liens de v

degré moyen $d^o = \frac{2m}{n}$



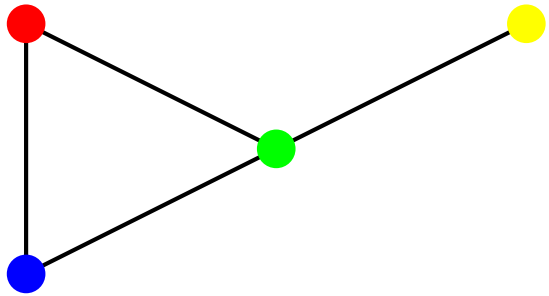
degrés : **2**, **2**, **3**, **1** ; degré moyen 2

Propriétés de graphes

degré d'un nœud v, $d^o(v)$

= nombre de liens de v

degré moyen $d^o = \frac{2m}{n}$

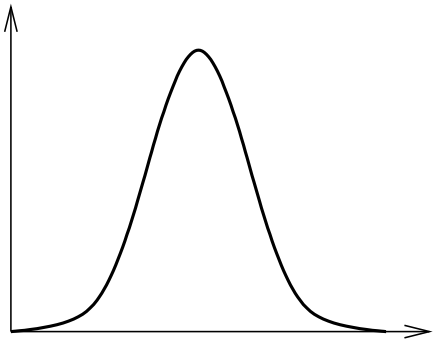


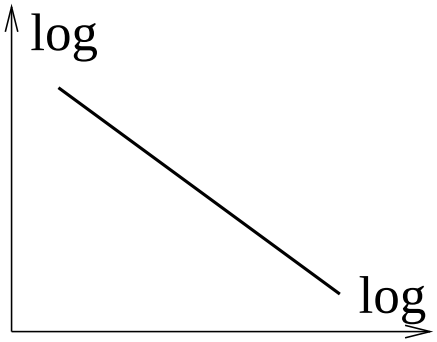
degrés : **2**, **2**, **3**, **1** ; degré moyen 2

distribution : $1 \to 1$, $2 \to 2$, $3 \to 1$

Un mot sur les distributions

deux grandes catégories



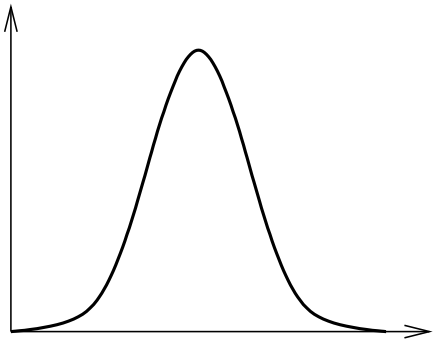


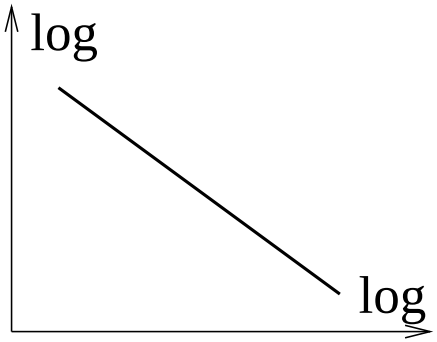
décroissance exponentielle	vs	décroissance polynômiale
Poisson, Gaussienne, ...	vs	Zipf, loi de puissance, ...
homogène	**vs**	**hétérogène**
tailles	vs	salaires
moyennne	vs	exposant

des conséquences importantes

Un mot sur les distributions

deux grandes catégories



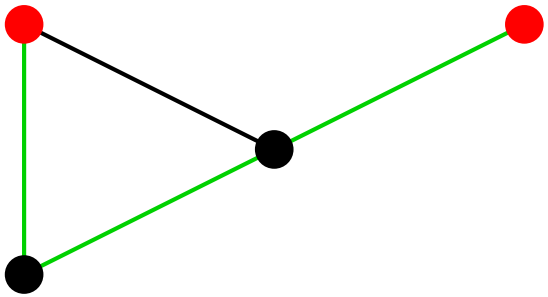


décroissance exponentielle	vs	décroissance polynômiale
Poisson, Gaussienne, ...	vs	Zipf, loi de puissance, ...
homogène	**vs**	**hétérogène**
tailles	vs	salaires
moyennne	vs	exposant

des conséquences importantes

Propriétés de graphes

chemin de *u* à *v* = suite de liens u...v

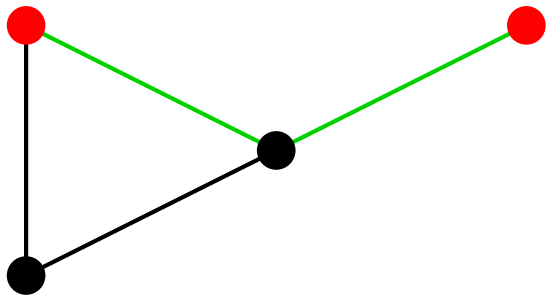


un chemin de longueur 3

Propriétés de graphes

chemin de u à v = suite de liens u...v

distance $d(u, v)$ = longueur d'*un* plus court chemin



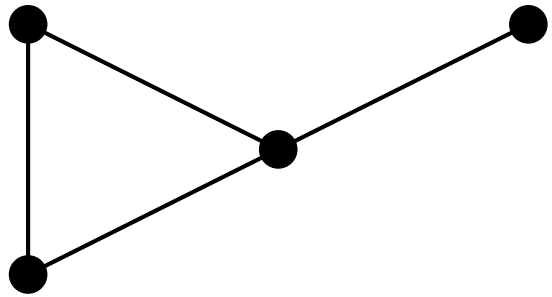
un plus court chemin ; longueur 2 $\Rightarrow$ distance = 2

Propriétés de graphes

chemin de u à v = suite de liens u...v

distance $d(u,v)$ = longueur d'*un* plus court chemin

diamètre Δ = plus grande distance

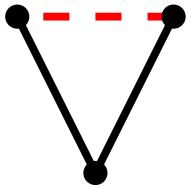


diamètre = 2

Propriétés de graphes

coefficient de clustering $\mathrm{cc}(v)$

= probabilité que deux voisins de v soient reliés

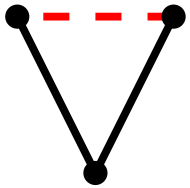


= densité locale

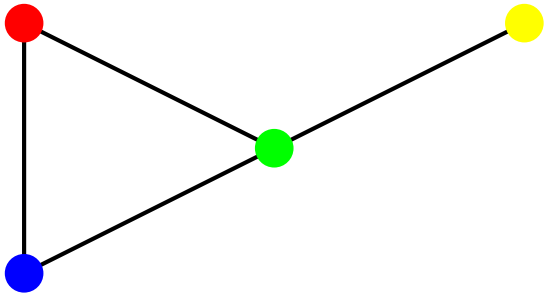
Propriétés de graphes

coefficient de clustering $\mathrm{cc}(v)$

= probabilité que deux voisins de v soient reliés



= densité locale



coefficients de clustering : 1, 1, $\frac{1}{3}$, indéfini

Graphes de terrain

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

propriétés en commun :

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

non-trivial ?

Graphes de terrain

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

propriétés en commun :

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

non-trivial ?

Graphes de terrain

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

propriétés en commun :

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

non-trivial ?

Comparaison à l'aléatoire

graphe aléatoire (Erdös, Rényi)
= choisi uniformément dans l'ensemble

graphe type, représentatif

des propriétés significatives,
d'autres moins

Comparaison à l'aléatoire

graphe aléatoire (Erdös, Rényi)
= choisi uniformément dans l'ensemble

graphe type, représentatif

	de terrain	aléatoire
densité	faible	faible
distances	faibles	faibles
degrés	hétérogènes	homogènes
clustering	fort	faible

des propriétés significatives,
d'autres moins

Comparaison à l'aléatoire

graphe aléatoire (Erdös, Rényi)
= choisi uniformément dans l'ensemble

graphe type, représentatif

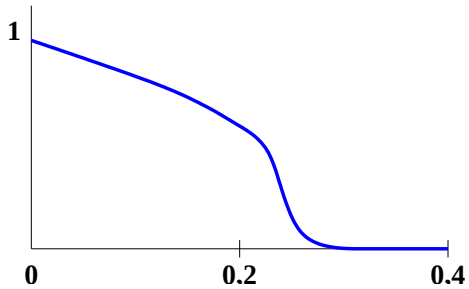
	de terrain	aléatoire	degrés prescrits
densité	faible	faible	faible
distances	faibles	faibles	faibles
degrés	hétérogènes	homogènes	hétérogènes
clustering	fort	faible	faible

des propriétés significatives,
d'autres moins

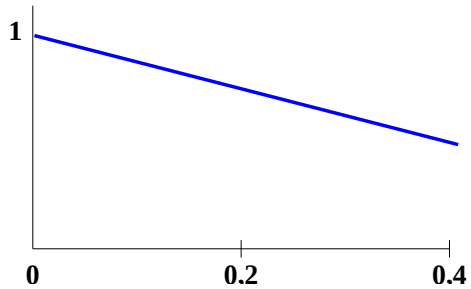
Conséquence : robustesse

pannes = suppression de nœuds au hasard.

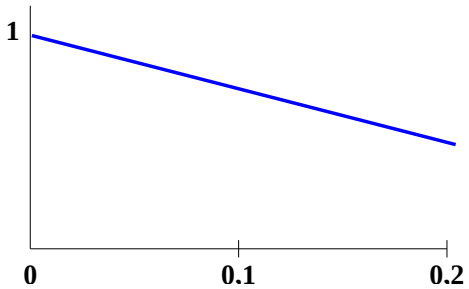
attaque = suppression de nœuds choisis (stratégie).



aléatoire



loi de puissance

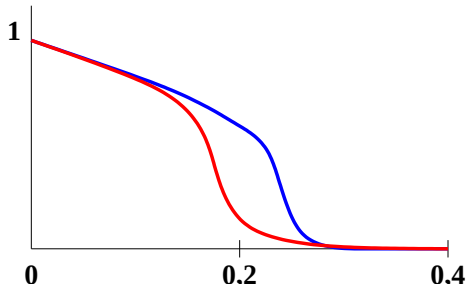


Internet

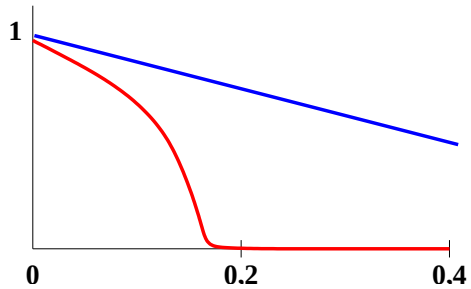
Conséquence : robustesse

pannes = suppression de nœuds au hasard.

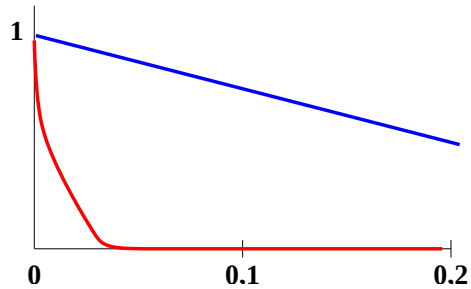
attaque = suppression de nœuds choisis (stratégie).



aléatoire



loi de puissance

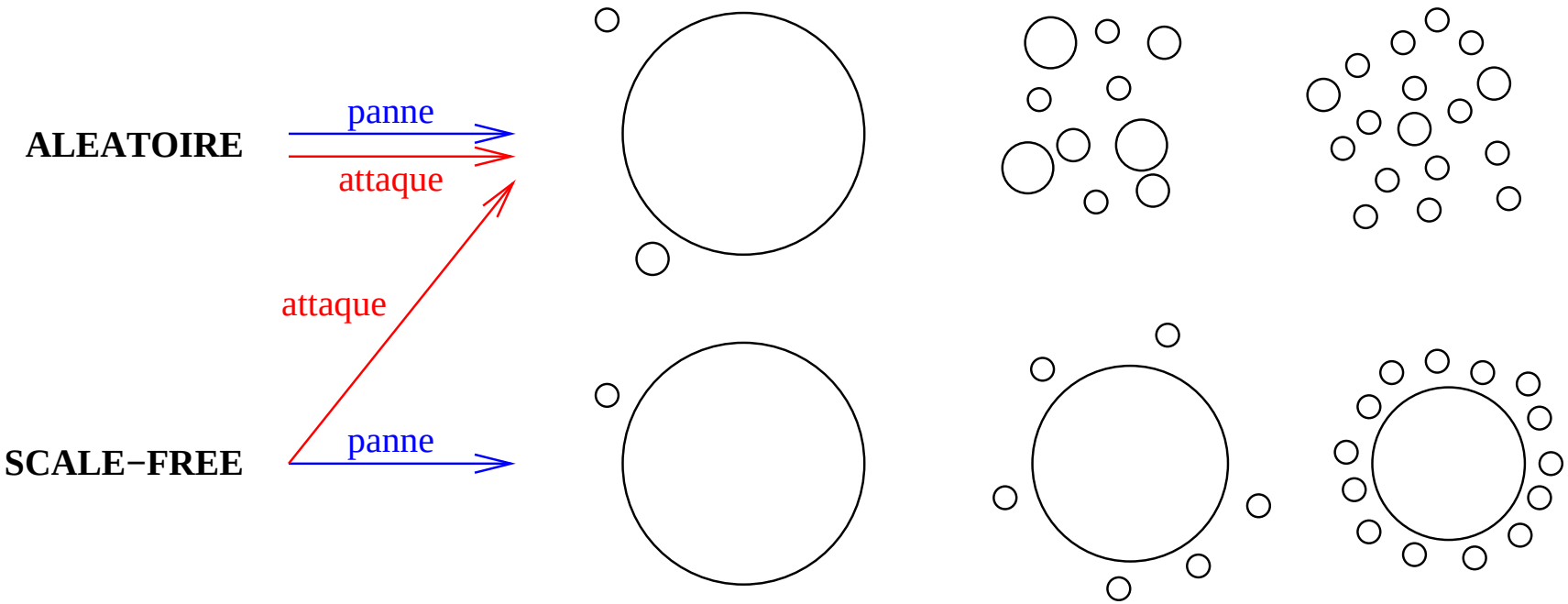


Internet

Conséquence : robustesse

pannes = suppression de nœuds au hasard.

attaque = suppression de nœuds choisis (stratégie).



Conséquence : robustesse

pannes = suppression de nœuds au hasard.

attaque = suppression de nœuds choisis (stratégie).

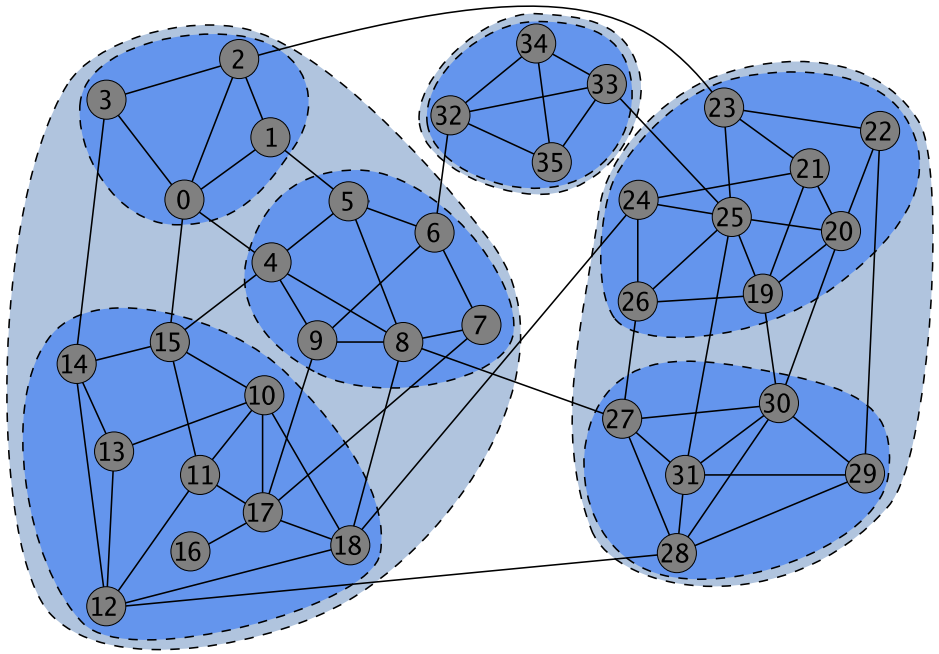
↪ **Très résistant aux pannes.**

↪ **Sensible aux attaques.**

↪ **Importance de la topologie sous-jacente.**

Conséquence : communautés

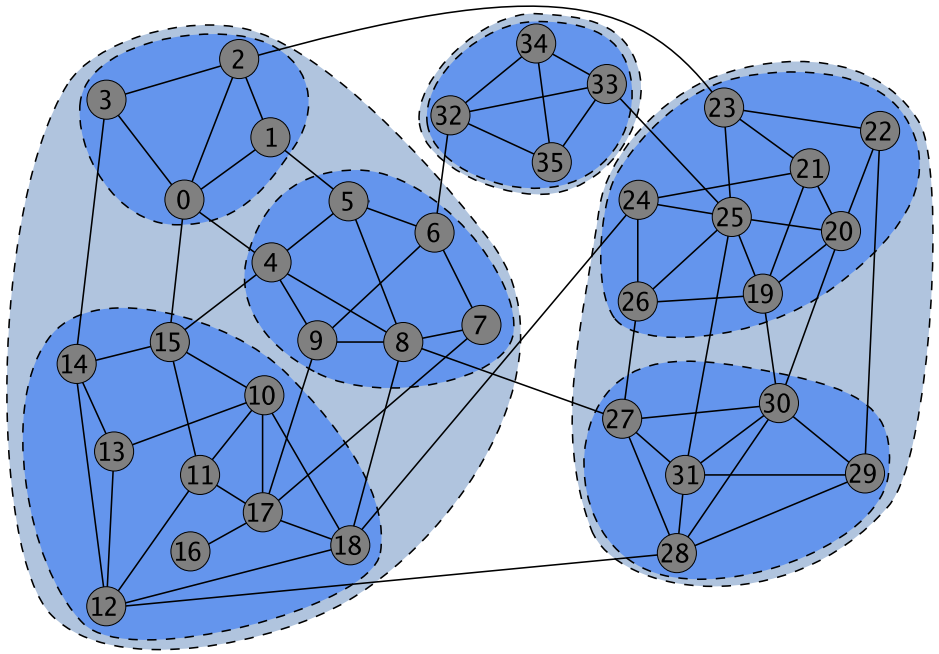
zones denses faiblement interconnectées



structure hiérarchique
recouvrements
calcul ?

Conséquence : communautés

zones denses faiblement interconnectées



structure hiérarchique
recouvrements
calcul ?

Graphes de terrain

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

propriétés en commun :

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

et des problématiques communes

- **mesure et métrologie**
- **modélisation**
- **analyse**
- **algorithmique**

Graphes de terrain

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

propriétés en commun :

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

et des problématiques communes

- **mesure et métrologie**
- **modélisation**
- **analyse**
- **algorithmique**

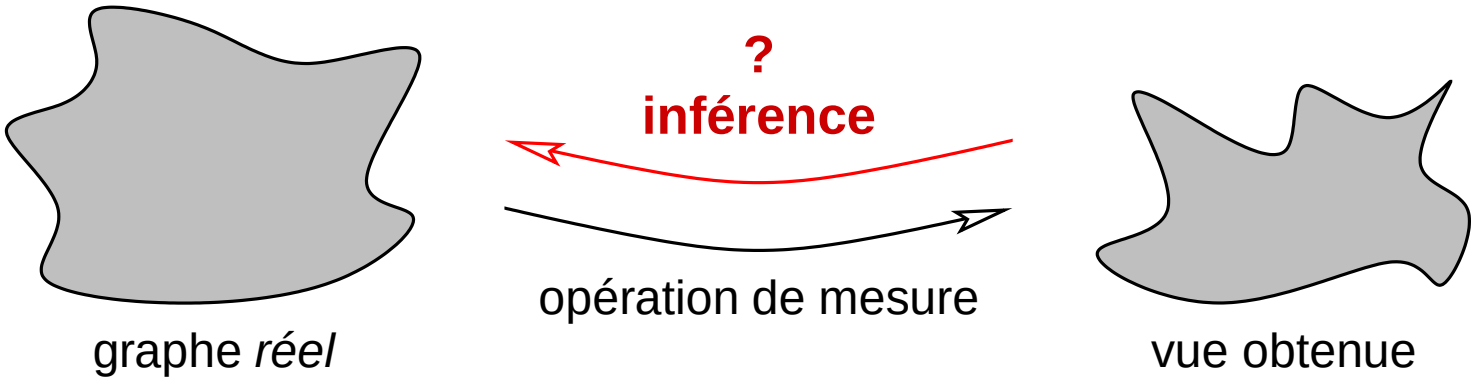
Mesure

graphe *réel* — vue obtenue

que peut on dire sur l'objet réel à partir de la mesure ?

impact sur les propriétés observées ?
impact des propriétés sur la vue ?
mesures ciblant certaines propriétés ?
...

Mesure et Métrologie



que peut on dire sur l'objet réel à partir de la mesure ?

impact sur les propriétés observées ?
impact des propriétés sur la vue ?
mesures ciblant certaines propriétés ?
...

Analyse

décrire

extraire de l'information pertinente

statistiques

densité
degrés
densité locale
corrélations
...

structure

encore beaucoup de manques

Analyse

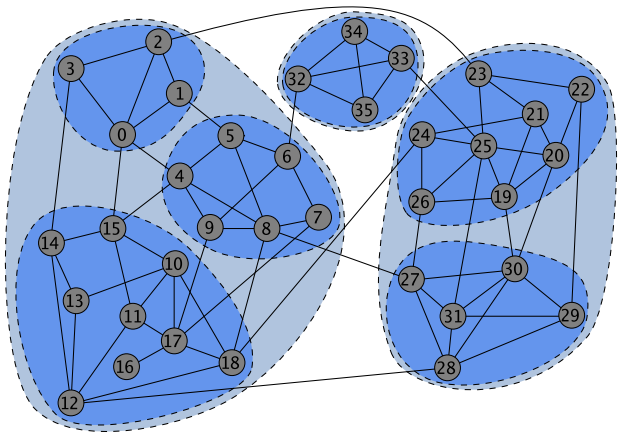
décrire

extraire de l'information pertinente

statistiques

densité
degrés
densité locale
corrélations
...

structure



encore beaucoup de manques

Modélisation

génération de graphes *réalistes*
(*i.e.* ayant les propriétés observées)

motivations : approches formelles, simulation, explication

état de l'art :

- taille, densité, distances : facile
- degrés : consensus, modèles bien maîtrisés
- densité locale : pas de consensus

Modélisation

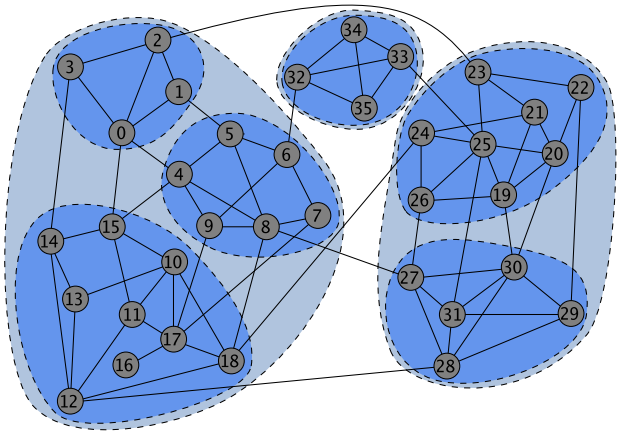
génération de graphes *réalistes*
(*i.e.* ayant les propriétés observées)

motivations : approches formelles, simulation, explication

état de l'art :

- taille, densité, distances : facile
- degrés : consensus, modèles bien maîtrisés
- densité locale : pas de consensus

Algorithmique

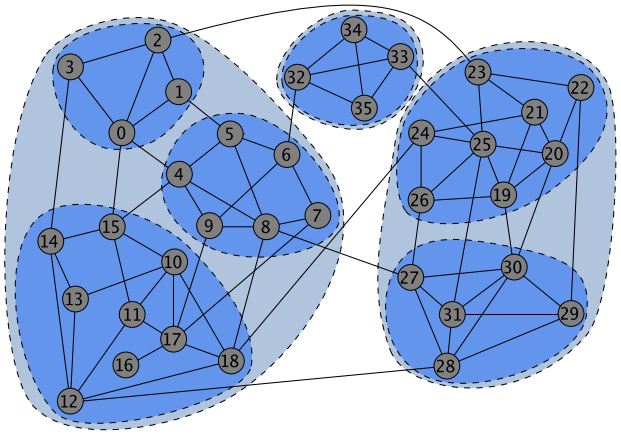


problèmes spécifiques
problèmes classiques à revisiter
restrictions en espace

- tirer parti des propriétés observées
- résultats approchés souvent suffisants
- algorithmique empirique / expérimentale

$\hookrightarrow$ **renouveau algorithmique**

Algorithmique



problèmes spécifiques
problèmes classiques à revisiter
restrictions en espace

- tirer parti des propriétés observées
- résultats approchés souvent suffisants
- algorithmique empirique / expérimentale

↪ **renouveau algorithmique**

Et la dynamique ?

(Tous) les graphes de terrain sont dynamiques

Ajout et suppression au fil du temps

- de nœuds
- de liens

Graphes de terrain

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

sciences sociales : collaboration, amitié, contacts sexuels, échanges, économie, etc

biologie : cerveau, gènes, protéines, écosystèmes, etc

linguistique : synonymie, co-occurrence, etc

transport : routier, aérien, électrique, etc

etc, etc

Prise en compte de la dynamique

Problématiques pour les graphes de terrain

- Mesure
- Analyse
- Modélisation
- Algorithmique

Prise en compte de la dynamique

Problématiques pour les graphes de terrain

- Mesure
- Analyse
- Modélisation
- Algorithmique

Dynamique $\longrightarrow$ nouvelles questions

But du cours

Illustrer les questions liées à l'étude de la dynamique

Pas encore de réponses !

S'appuyer sur deux cas concrets :

- topologie de l'internet
- trafic sur l'internet

Topologie physique de l'Internet

- Routeurs
- Câbles

Topologie physique de l'Internet

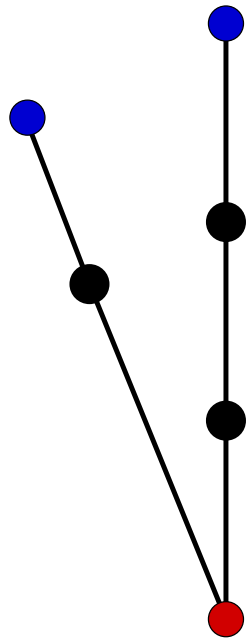
Mesure : Agrégation de chemins

Topologie physique de l'Internet

Mesure : Agrégation de chemins

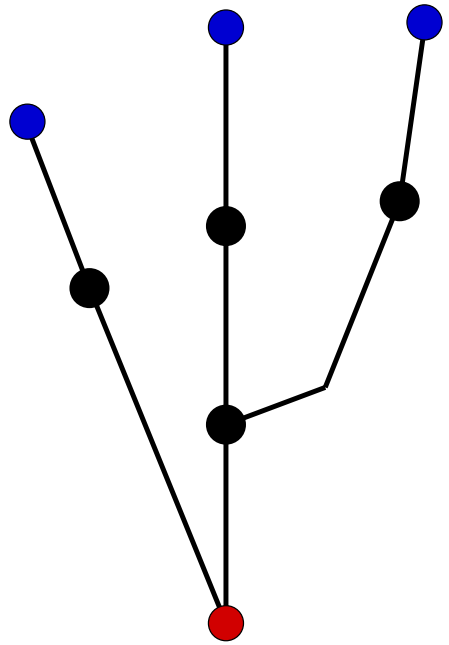
Topologie physique de l'Internet

Mesure : Agrégation de chemins



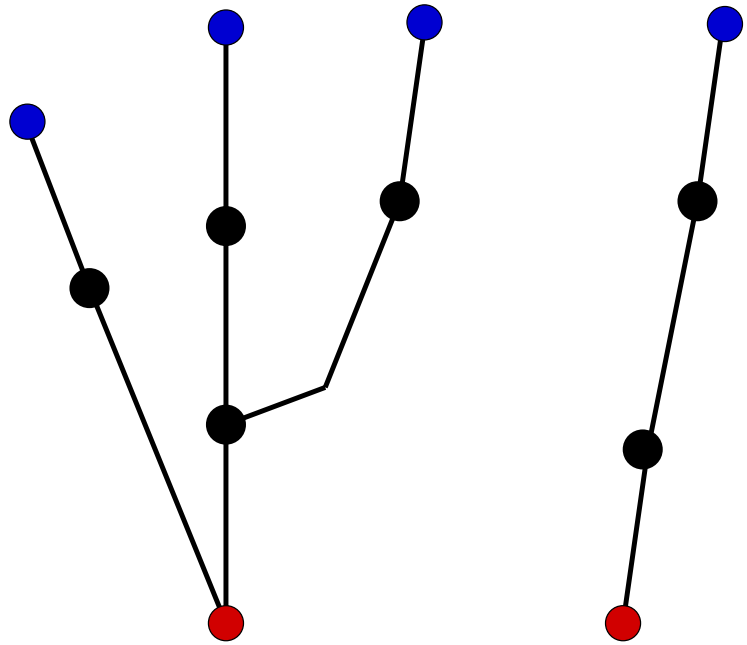
Topologie physique de l'Internet

Mesure : Agrégation de chemins



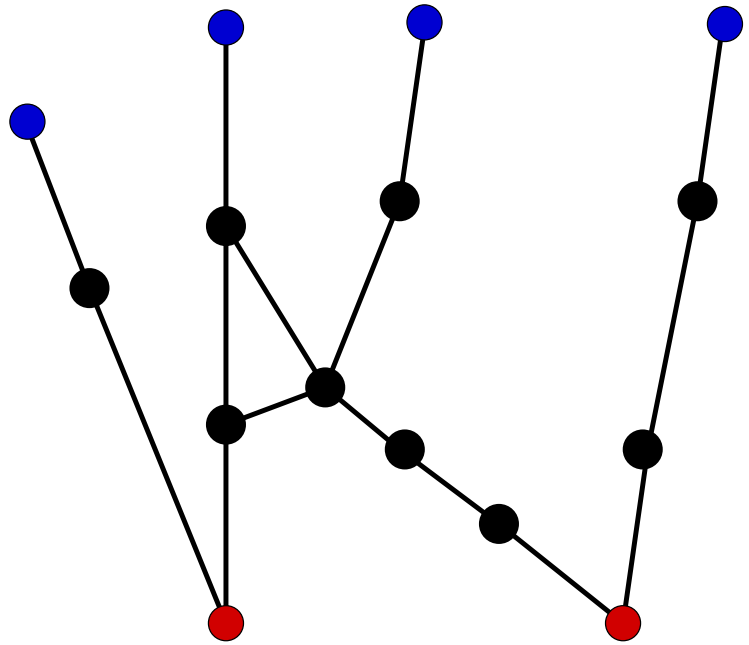
Topologie physique de l'Internet

Mesure : Agrégation de chemins



Topologie physique de l'Internet

Mesure : Agrégation de chemins



Topologie physique de l'Internet

Plus d'informations

Multiplier

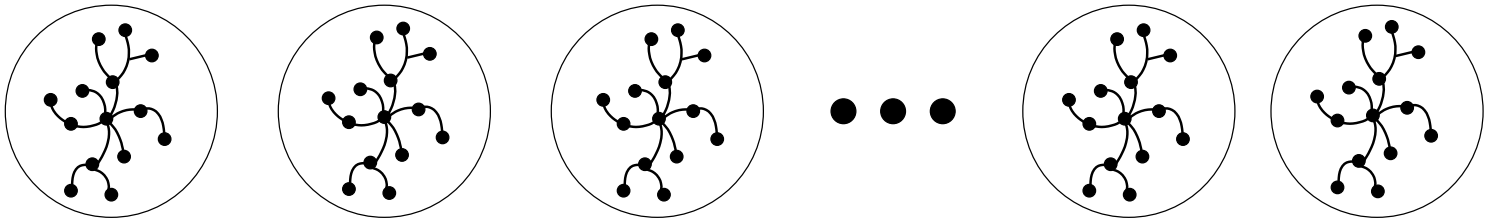
- Les destinations ($\sim$ 1 million)
- Les sources ($\sim$ quelques dizaines)

Processus coûteux et long
Dynamique ?

Notre approche

Vue égo-centrée

Chemins entre un moniteur et des destinations
Mesures périodiques



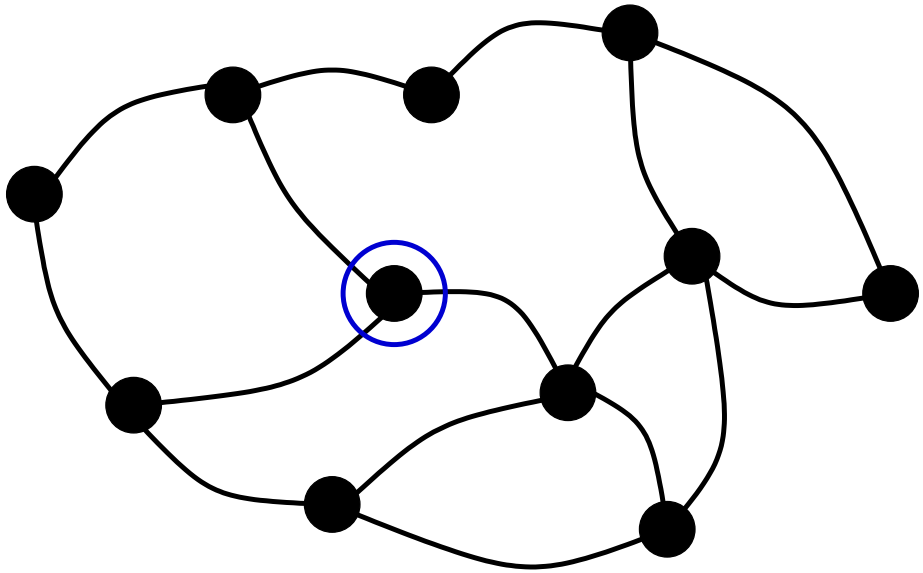
- $\sim$ 100 passes / jour
- plusieurs mois de mesure

Trafic sur l'internet

Qui communique avec qui ?

quelles machines

Capture sur un routeur

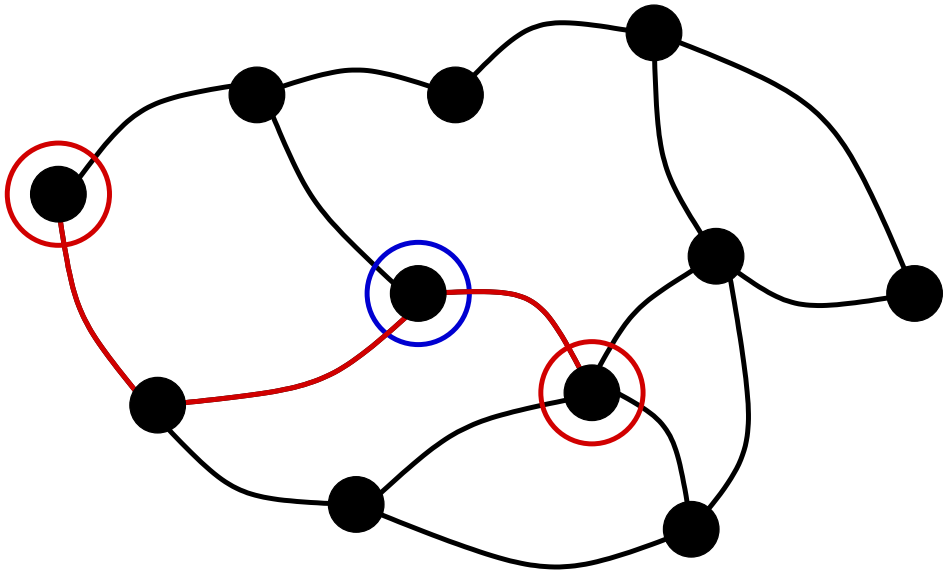


Trafic sur l'internet

Qui communique avec qui ?

quelles machines

Capture sur un routeur

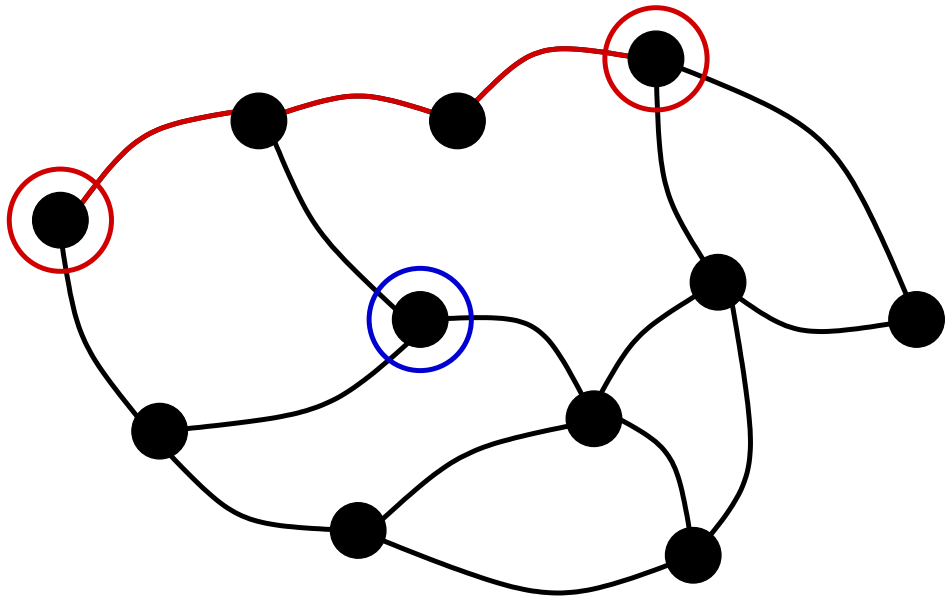


Trafic sur l'internet

Qui communique avec qui ?

quelles machines

Capture sur un routeur

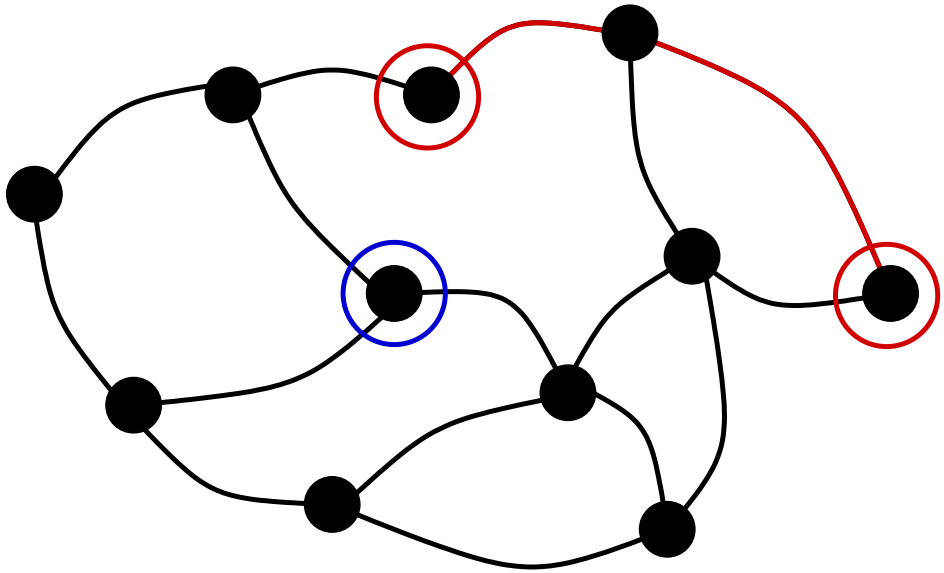


Trafic sur l'internet

Qui communique avec qui ?

quelles machines

Capture sur un routeur



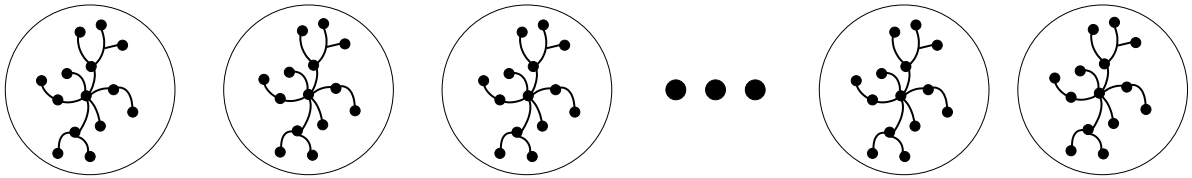
Trafic sur l'internet

Nos mesures

- Routeur à la sortie d'un campus
- Une semaine de mesure

Différence entre les deux cas

Internet : suite de graphes



Trafic : suite d'événements

```
t1 noeud1 noeud2
t2 noeud1 noeud2
t3 noeud1 noeud2
t4 noeud1 noeud2
.
.
.
tn noeud1 noeud2
```

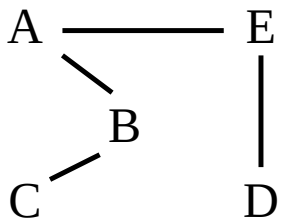
Événements : comment construire un graphe ?

1 A B
2 B C
2 D E
3 A E
4 A B
.
.
.
.
n X F

Agréger une fenêtre temporelle

Événements : comment construire un graphe ?

1 A B
2 B C
2 D E
3 A E
4 A B
.
.
.
.
n X F



Agréger une fenêtre temporelle

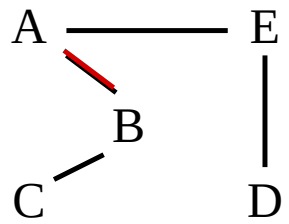
Événements : comment construire un graphe ?

1 A B
2 B C
2 D E
3 A E
4 A B
.
.
.
.
n X F

Agréger une fenêtre temporelle

Événements : comment construire un graphe ?

1 A B
2 B C
2 D E
3 A E
4 A B
.
.
.
.
n X F



Agréger une fenêtre temporelle

Taille de la fenêtre ?

Choix de la granularité ?

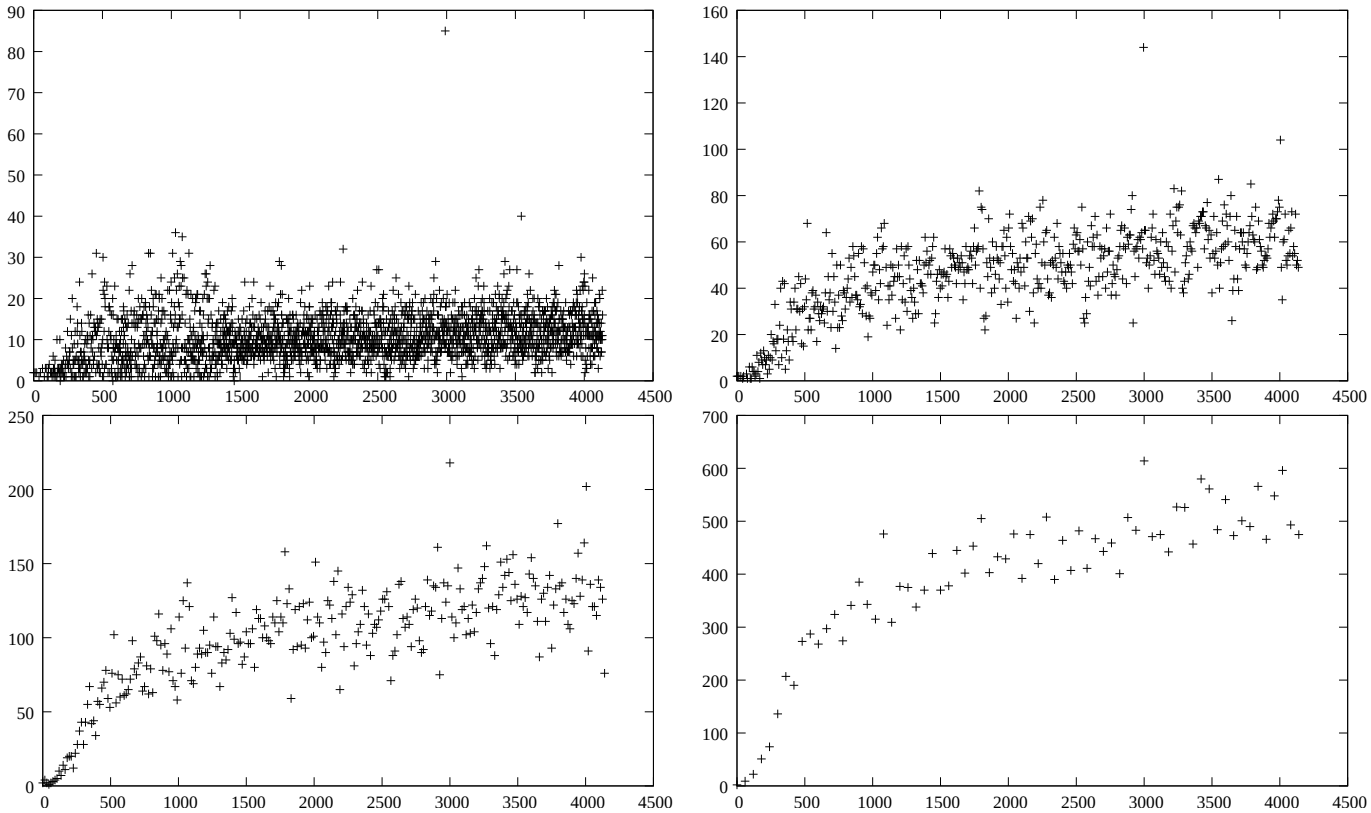
Taille de la fenêtre :

trop grande : perte d'informations
trop petite : graphes trops petits

Choix de la granularité ?

Choisir la plus petite période qui donne une régularité ?

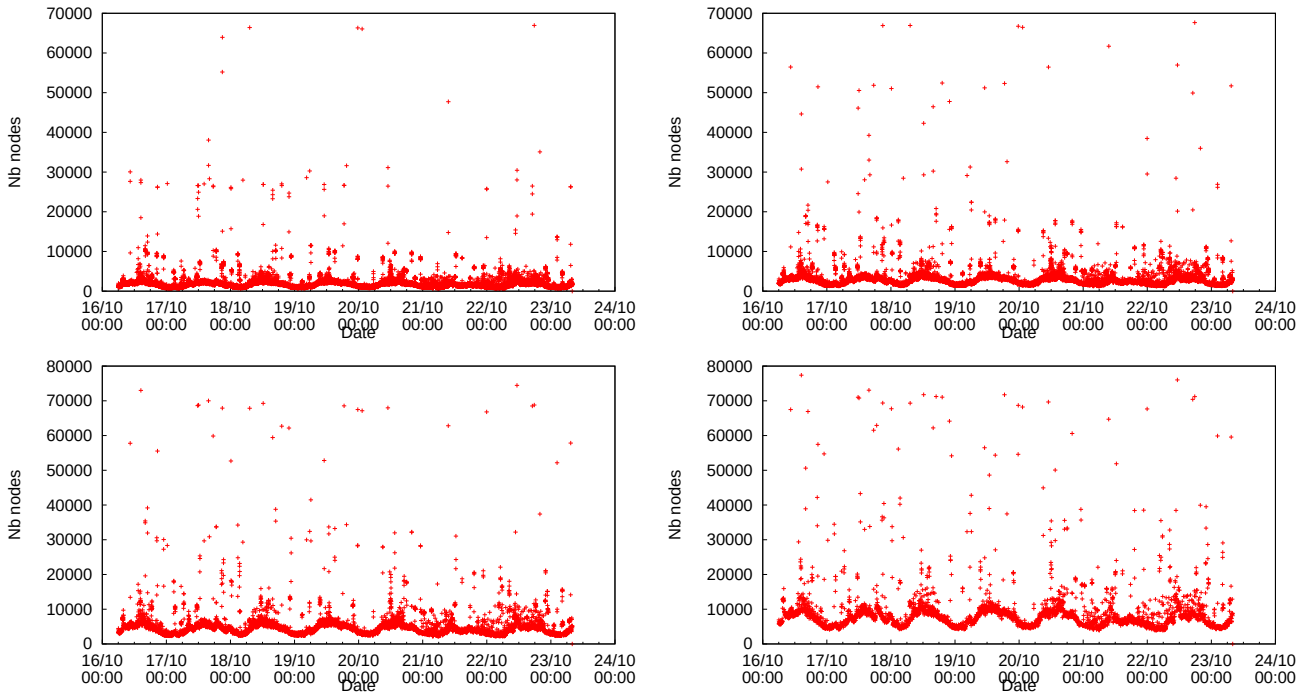
Marche (?) dans certains cas :



Choix de la granularité ?

Choisir la plus petite période qui donne une régularité ?

Pas dans d'autres (trafic internet)



Analyse de la dynamique

But : Comprendre, décrire

Détection d'événements

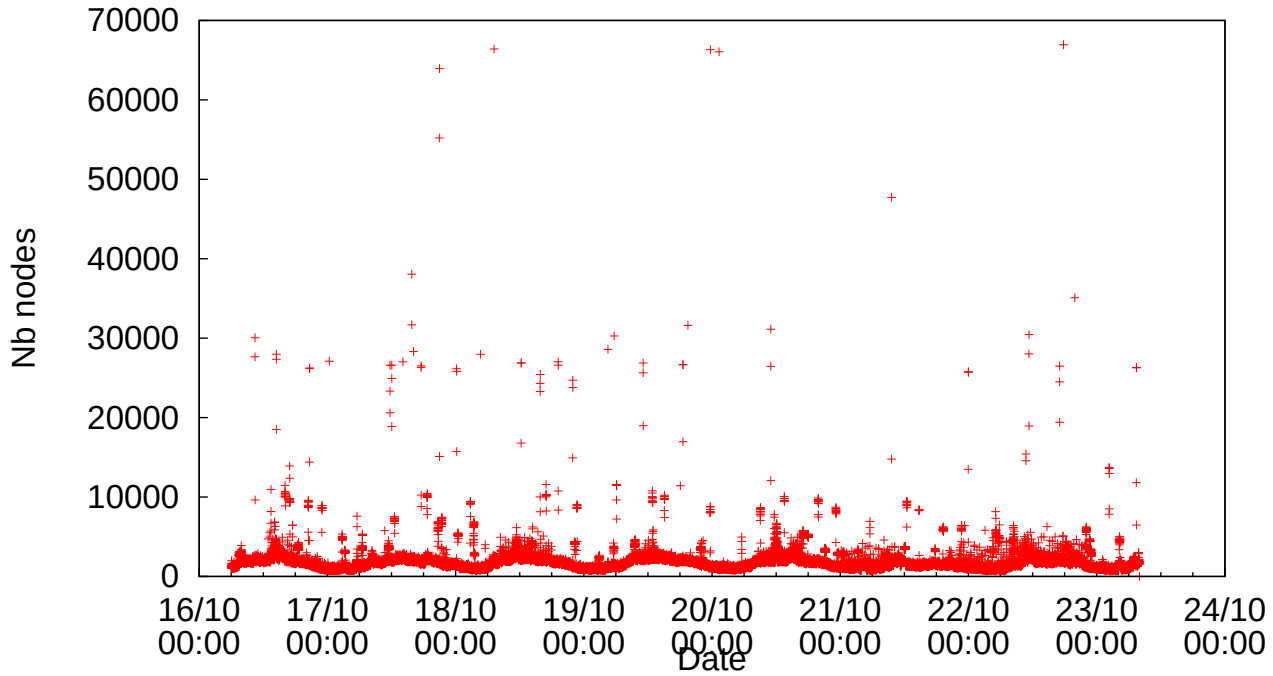
Paramètres pour les graphes statiques

Première approche :
Évolution de propriétés statiques au fil du temps

- nombre de nœuds
- nombre de liens
- degré moyen
- densité
- clustering
- ...

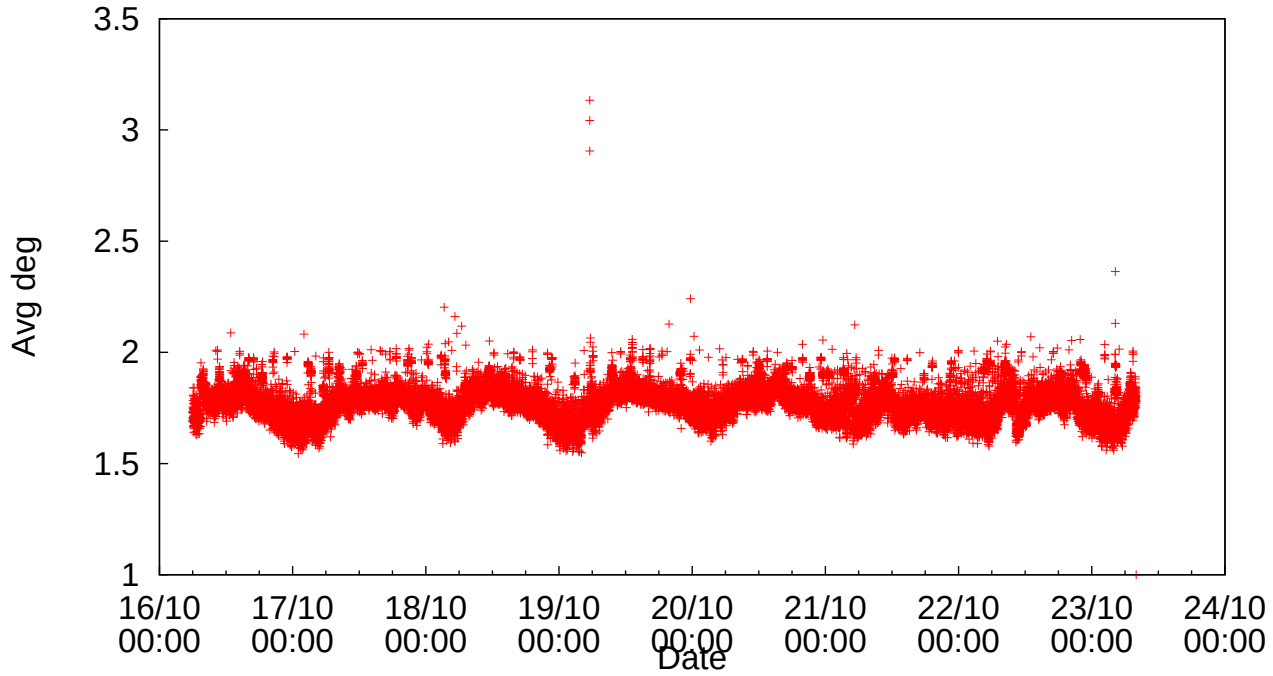
Paramètres pour les graphes statiques

Trafic internet, nombre de nœuds



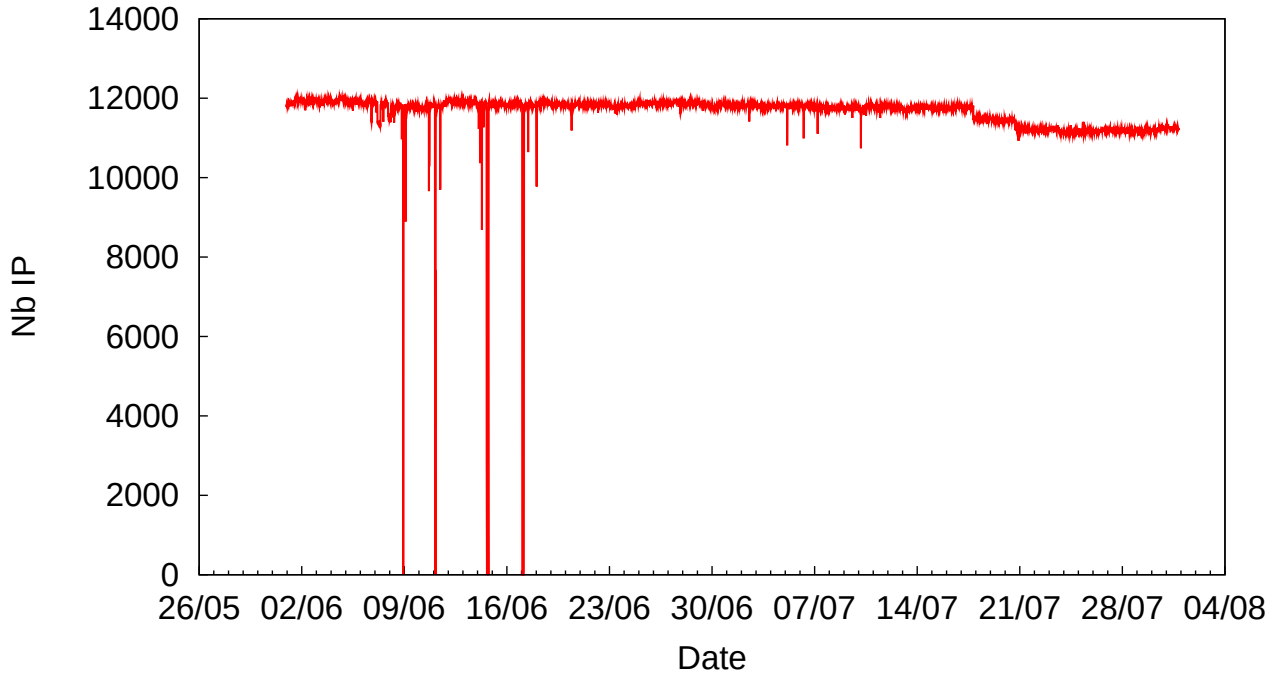
Paramètres pour les graphes statiques

Trafic internet, degré moyen



Paramètres pour les graphes statiques

Topologie internet, nombre de nœuds



Paramètres pour les graphes statiques

Observations

- Des comportements différents selon les graphes
- Détection de certaines tendances (effet jour/nuit)
- Événements (?)
- Traitement du signal ?

Autres propriétés ?

- Distribution des degrés ?
- Communautés ?

Distribution des degrés

- n signaux ?
- K-S test ?

Communautés

?

Graphes statiques : mauvaise approche ?

Évolution de paramètres statiques :

$\rightarrow$ peu d'informations sur l'évolution du graphe.

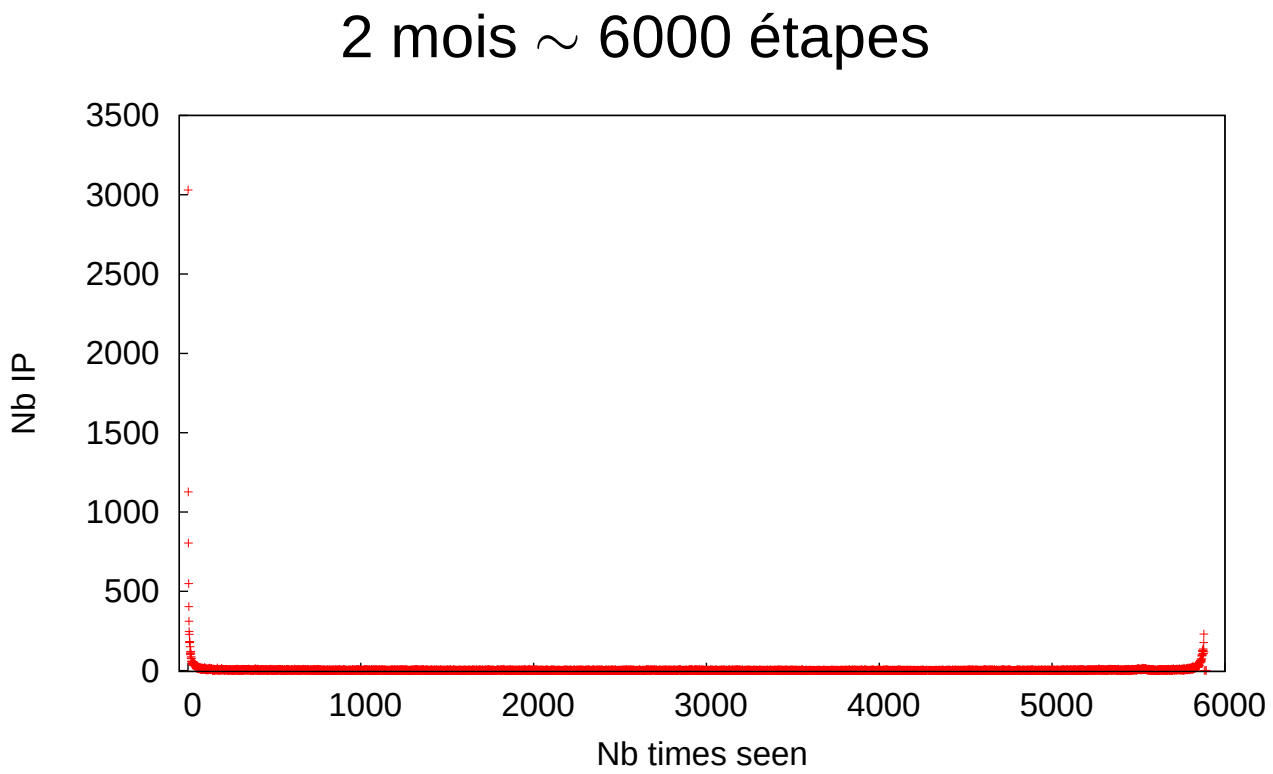
Ex : nombre de nœuds et liens constant $\rightarrow$?

Présence des nœuds – Topologie de l'internet

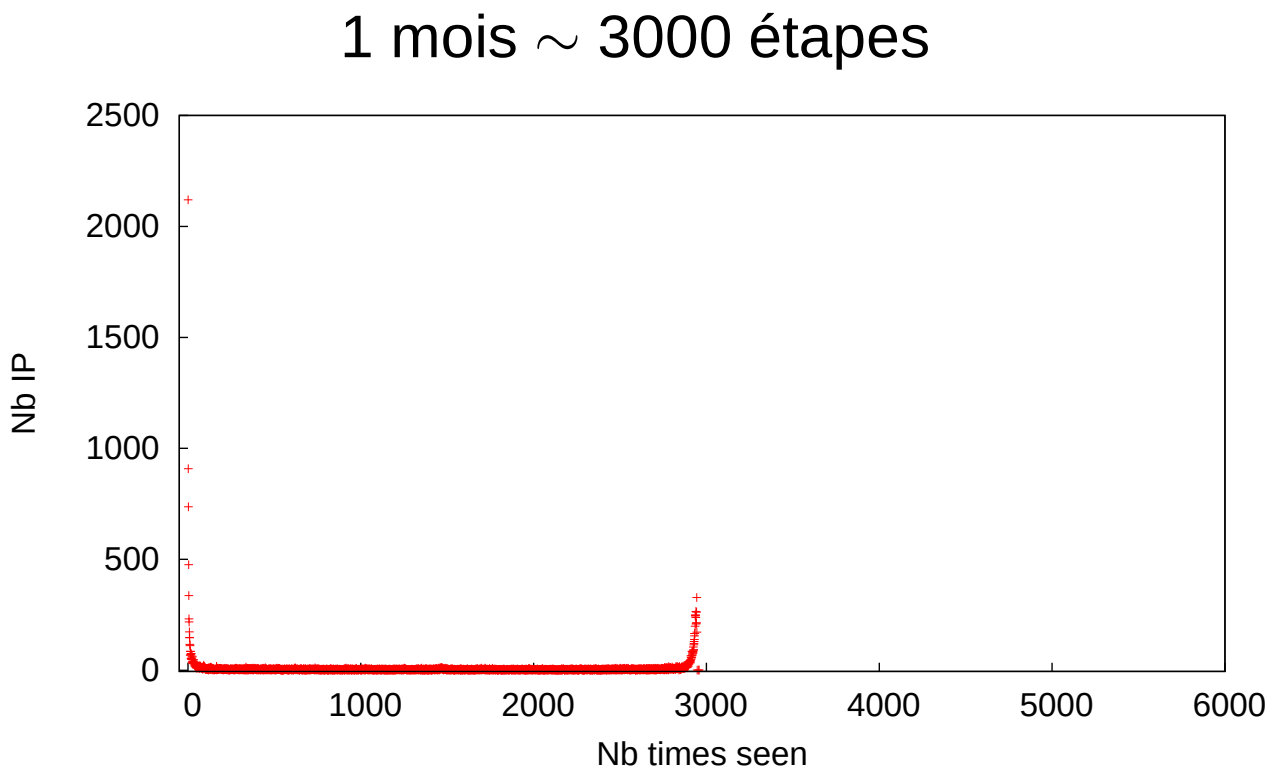
Distribution du nombre de présences

Tous les nœuds vus pendant la mesure ($\sim 29\,000$)
Pour chaque nœud : nombre d'étapes où on l'a vu
$\longrightarrow$ Distribution

Présence des nœuds – Topologie de l'internet

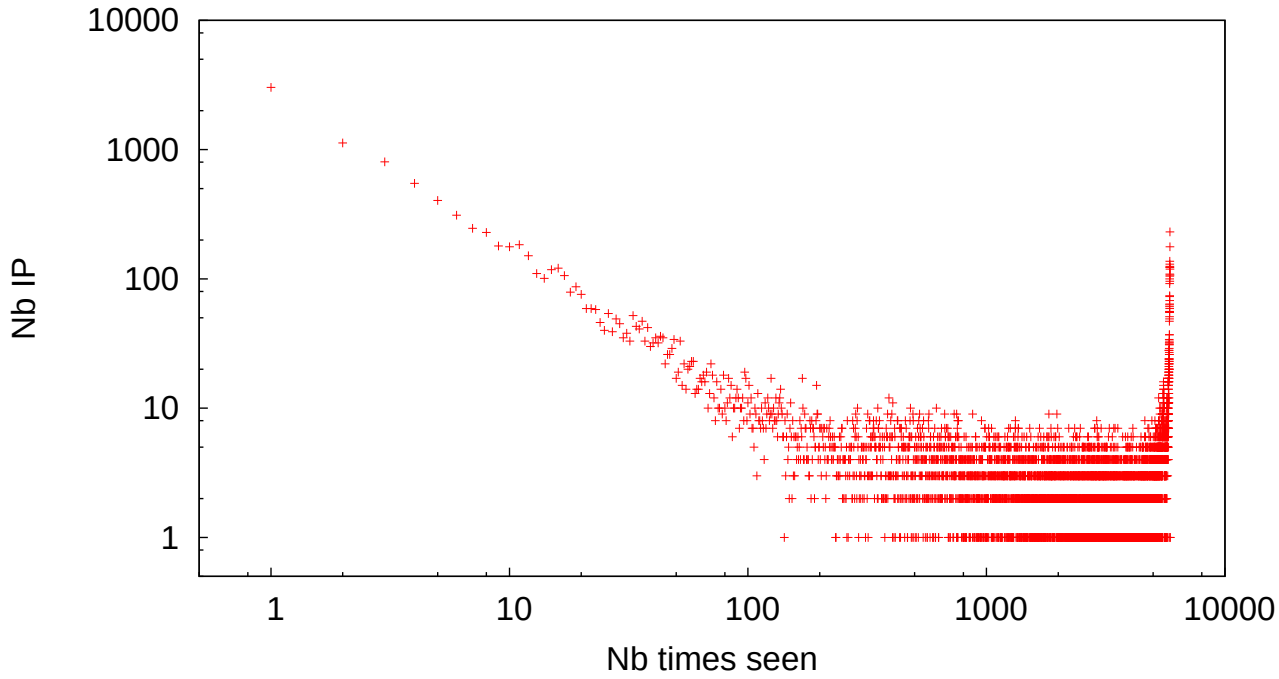


Présence des nœuds – Topologie de l'internet



Présence des nœuds – Topologie de l'internet

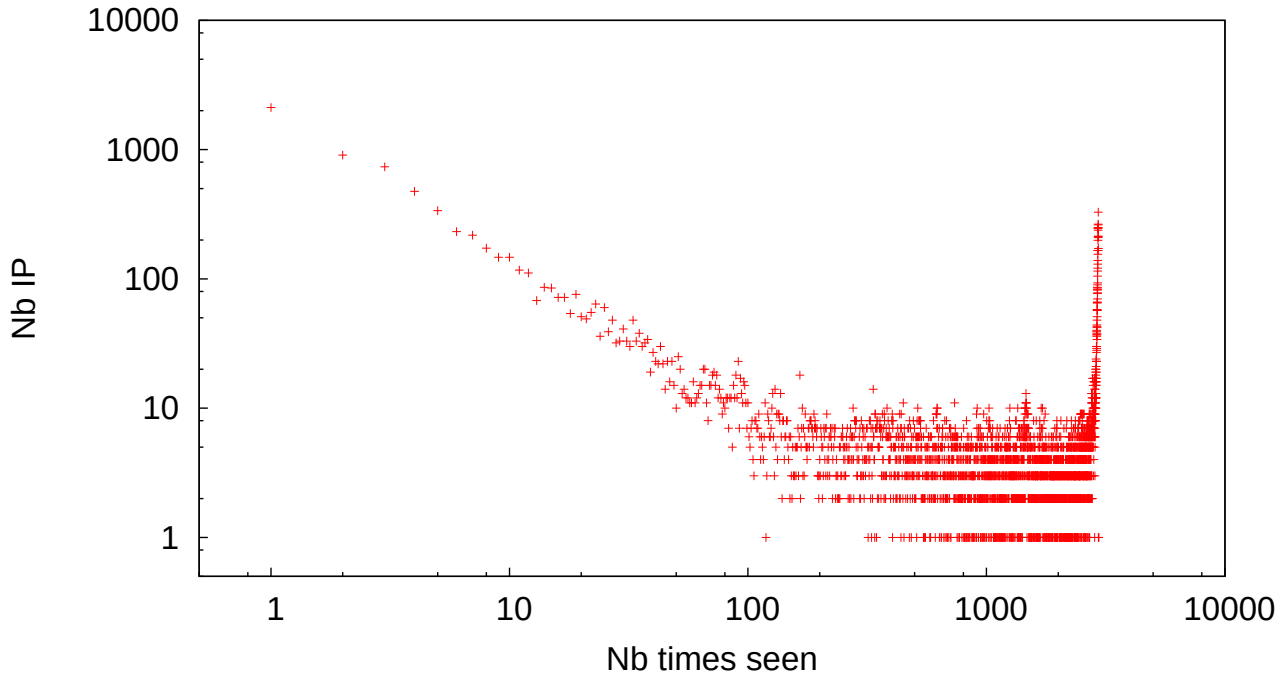
2 mois $\sim$ 6000 étapes — Échelle log-log



hétérogène $\rightarrow$ pas de valeur typique

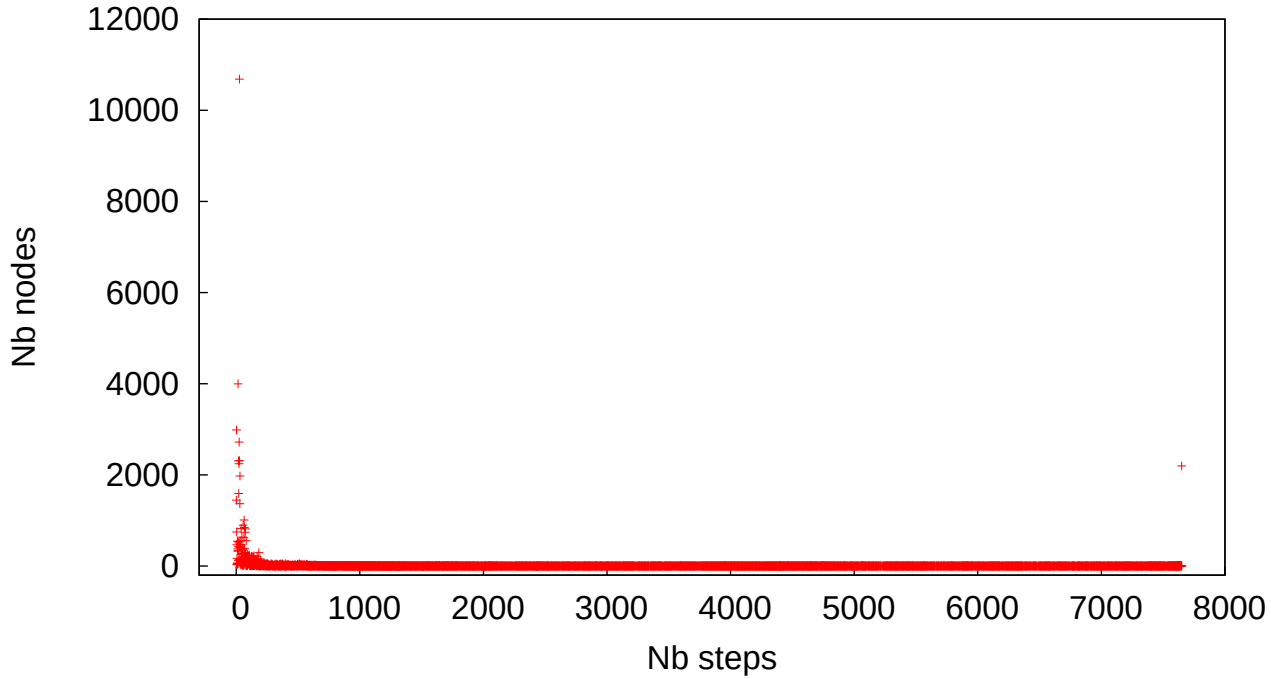
Présence des nœuds – Topologie de l'internet

1 mois $\sim$ 3000 étapes — Échelle log-log



hétérogène $\rightarrow$ pas de valeur typique

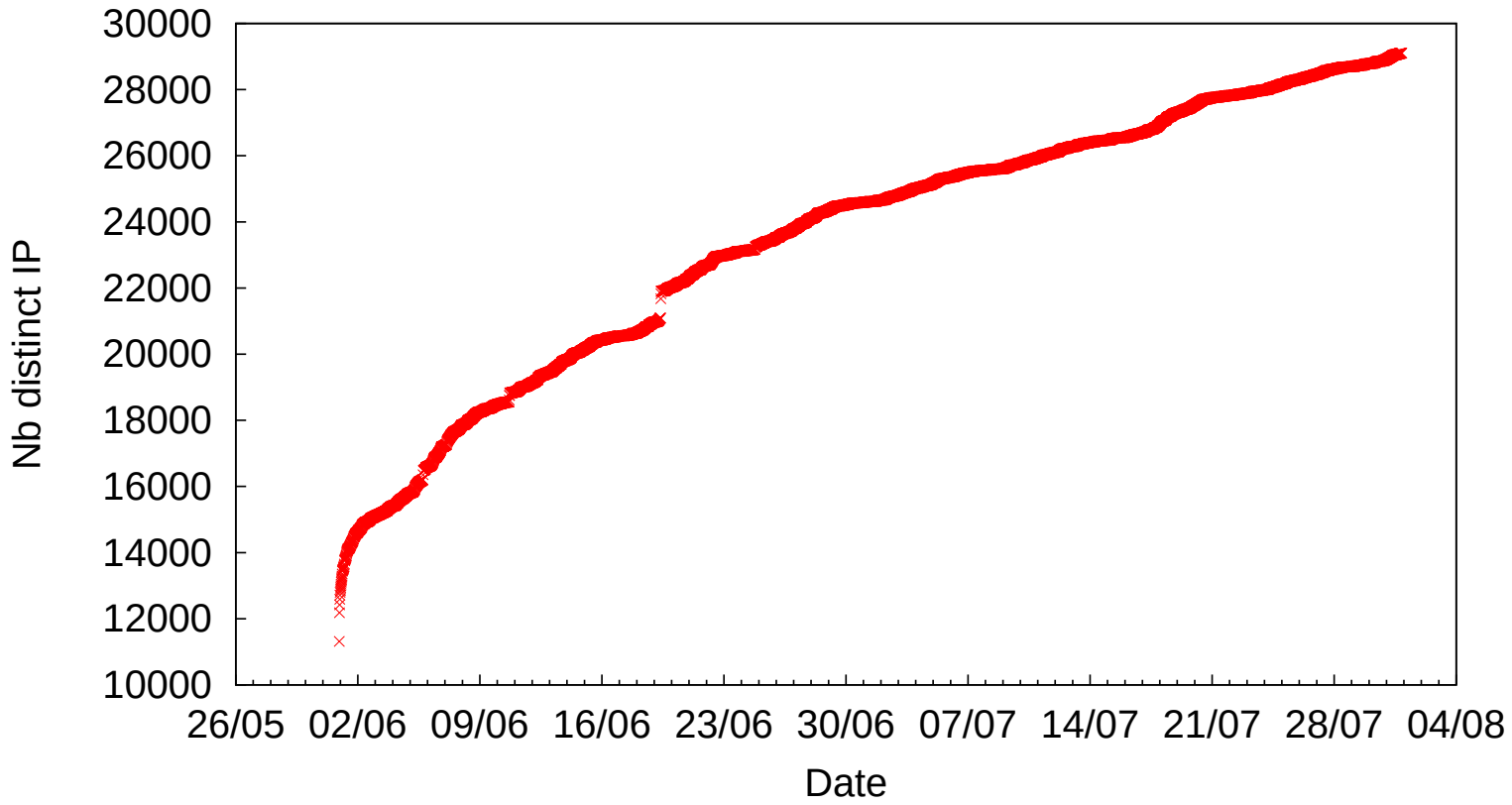
Présence des nœuds – Trafic internet



Similarité avec le cas précédent

Dynamique : stabilisation ?

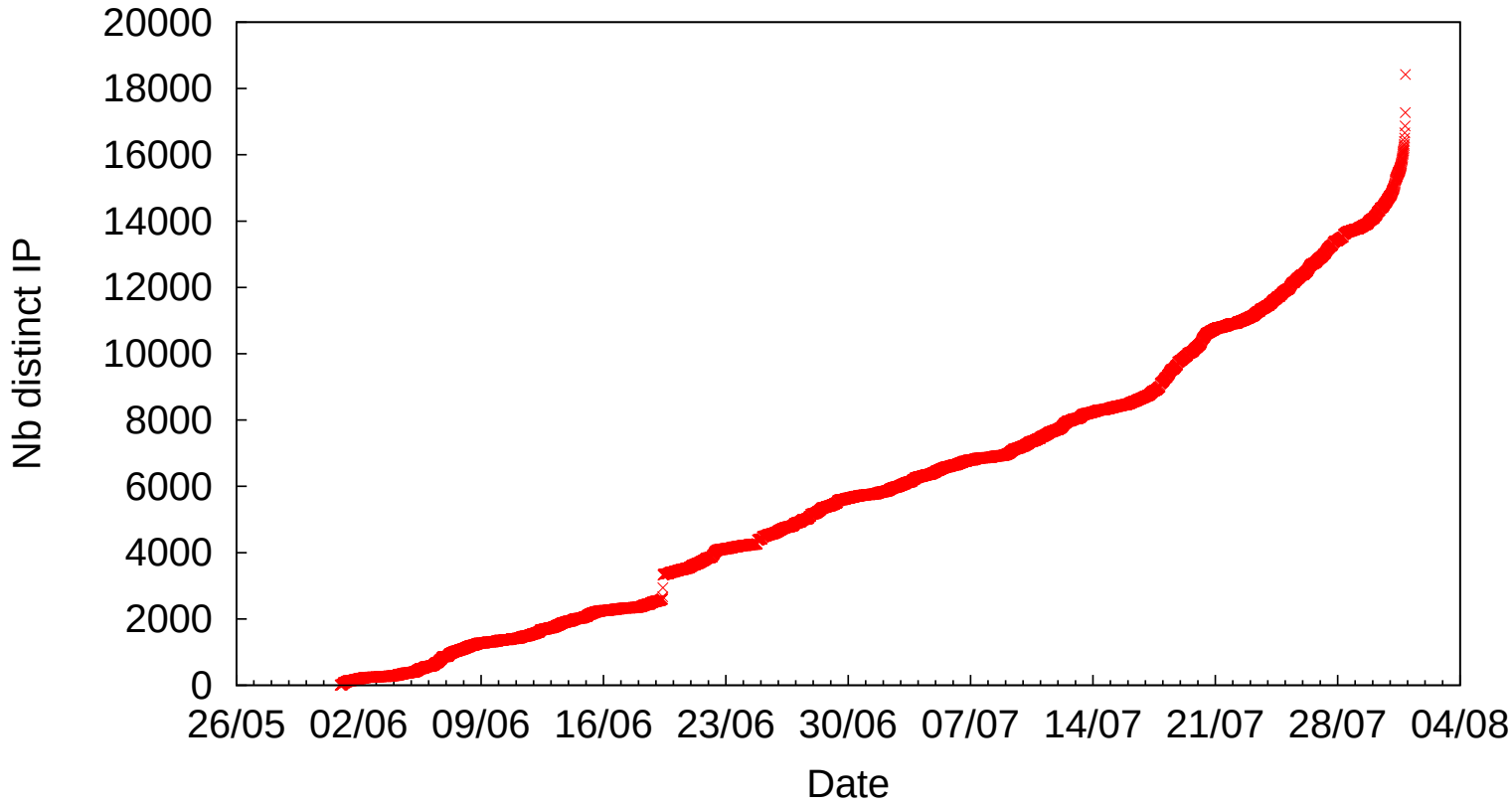
Nombre de nœuds distincts observés depuis le début
Topologie de l'internet



nouveaux nœuds observés en permanence

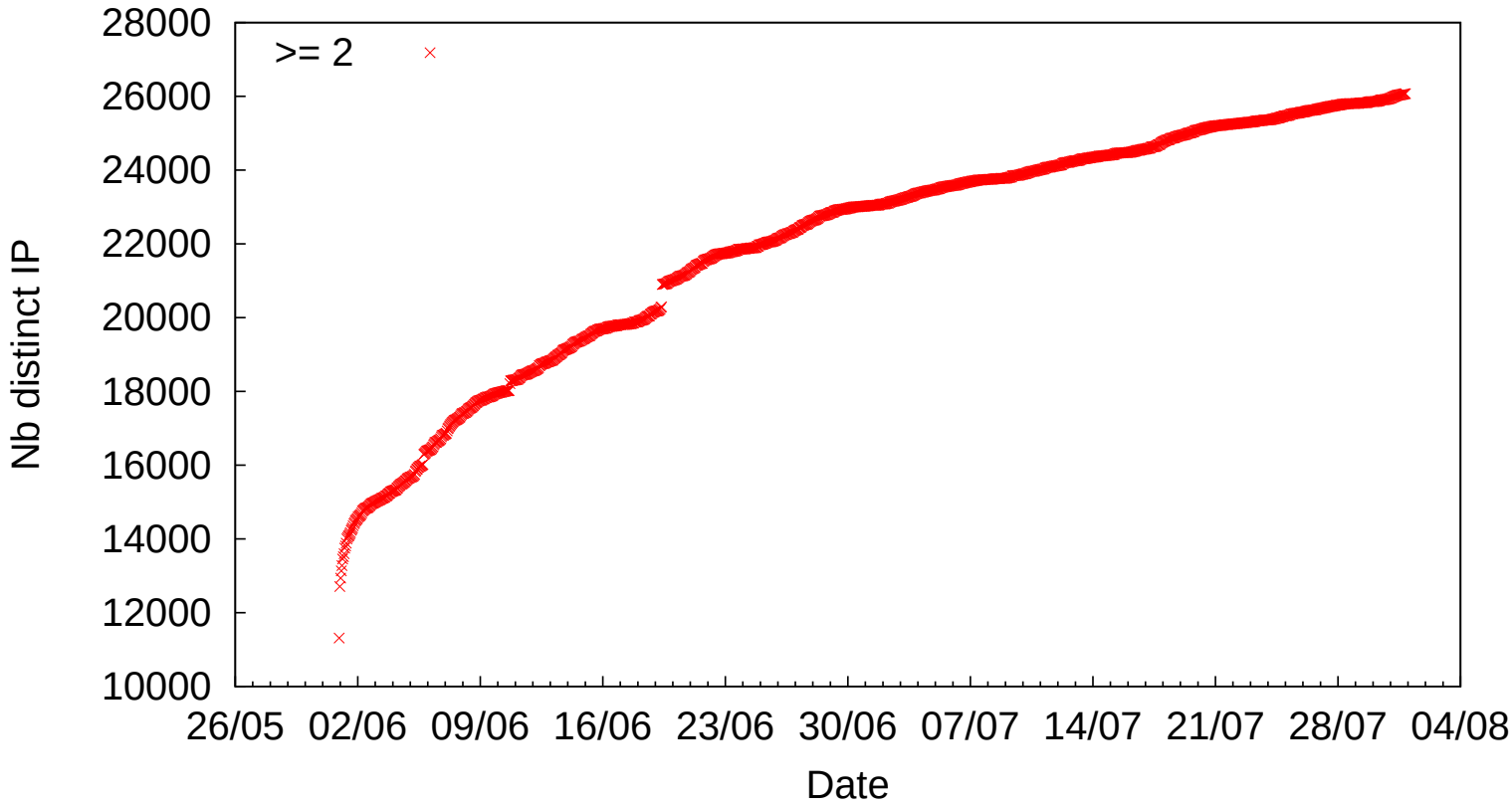
Disparitions

Nombre de nœuds distincts qu'on ne verra plus jamais
Topologie de l'internet

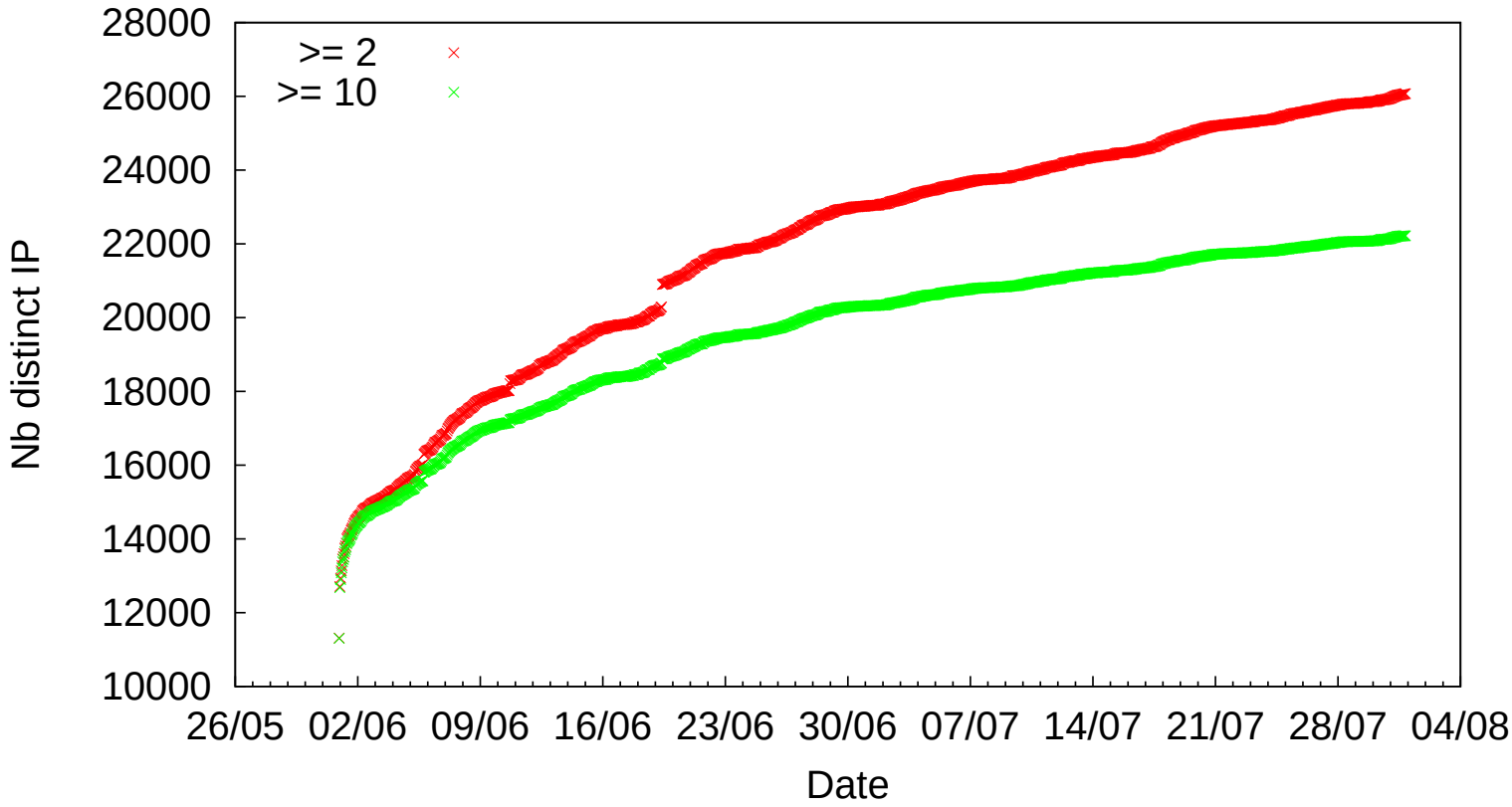


Renouvellement continu des nœuds observés

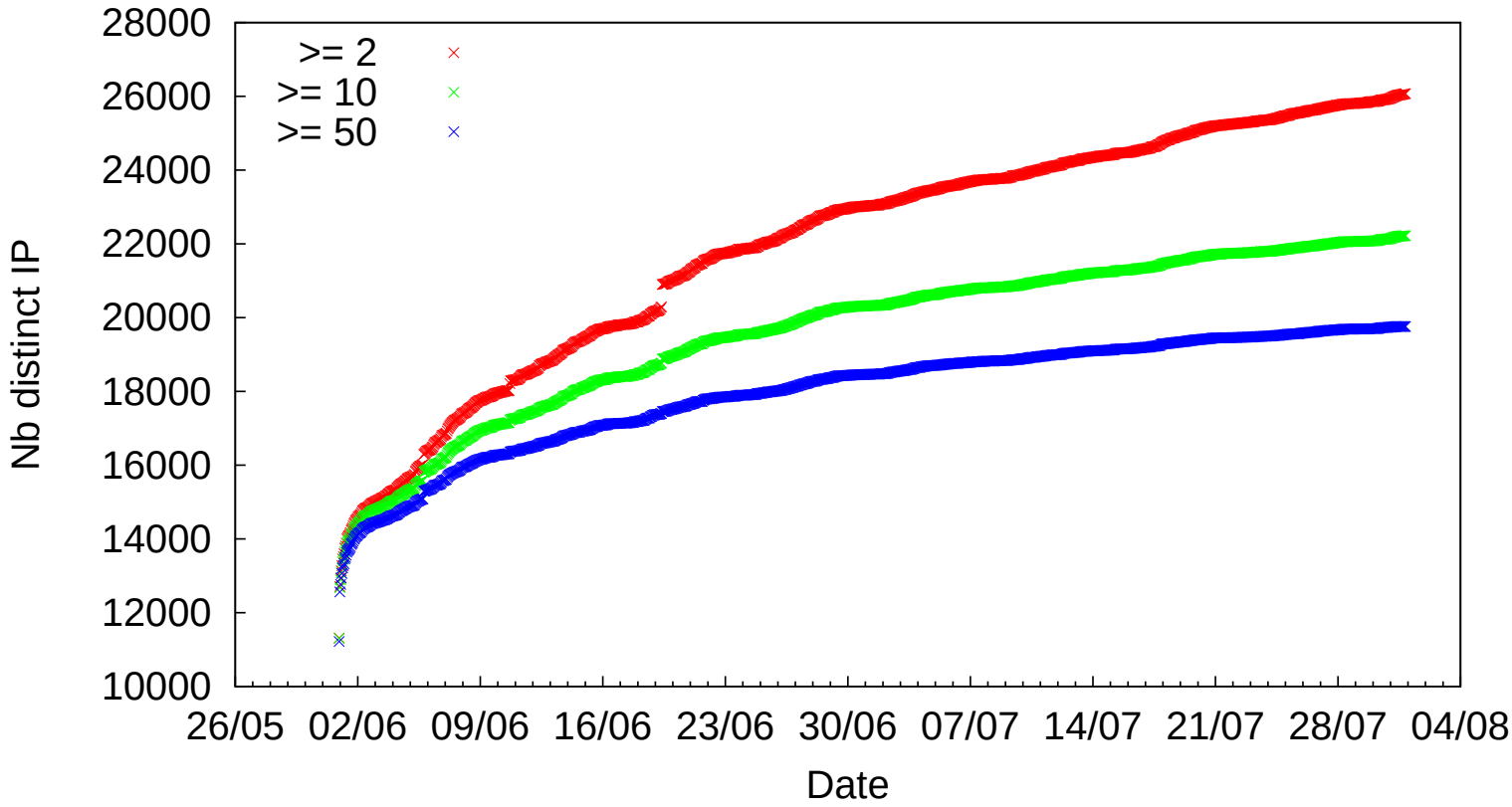
Influence des nœuds vus une seule fois ?



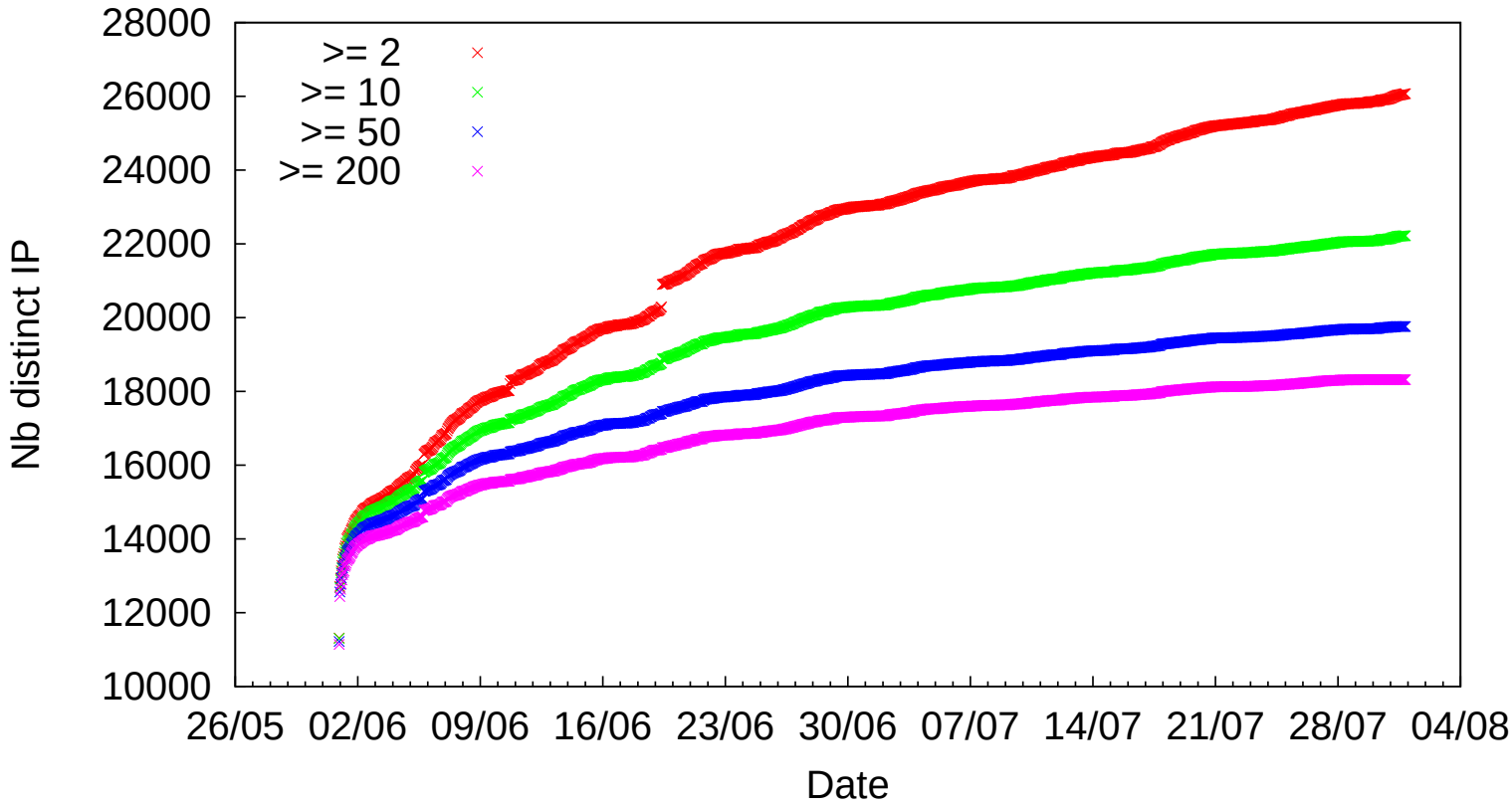
Influence des nœuds vus une seule fois ?



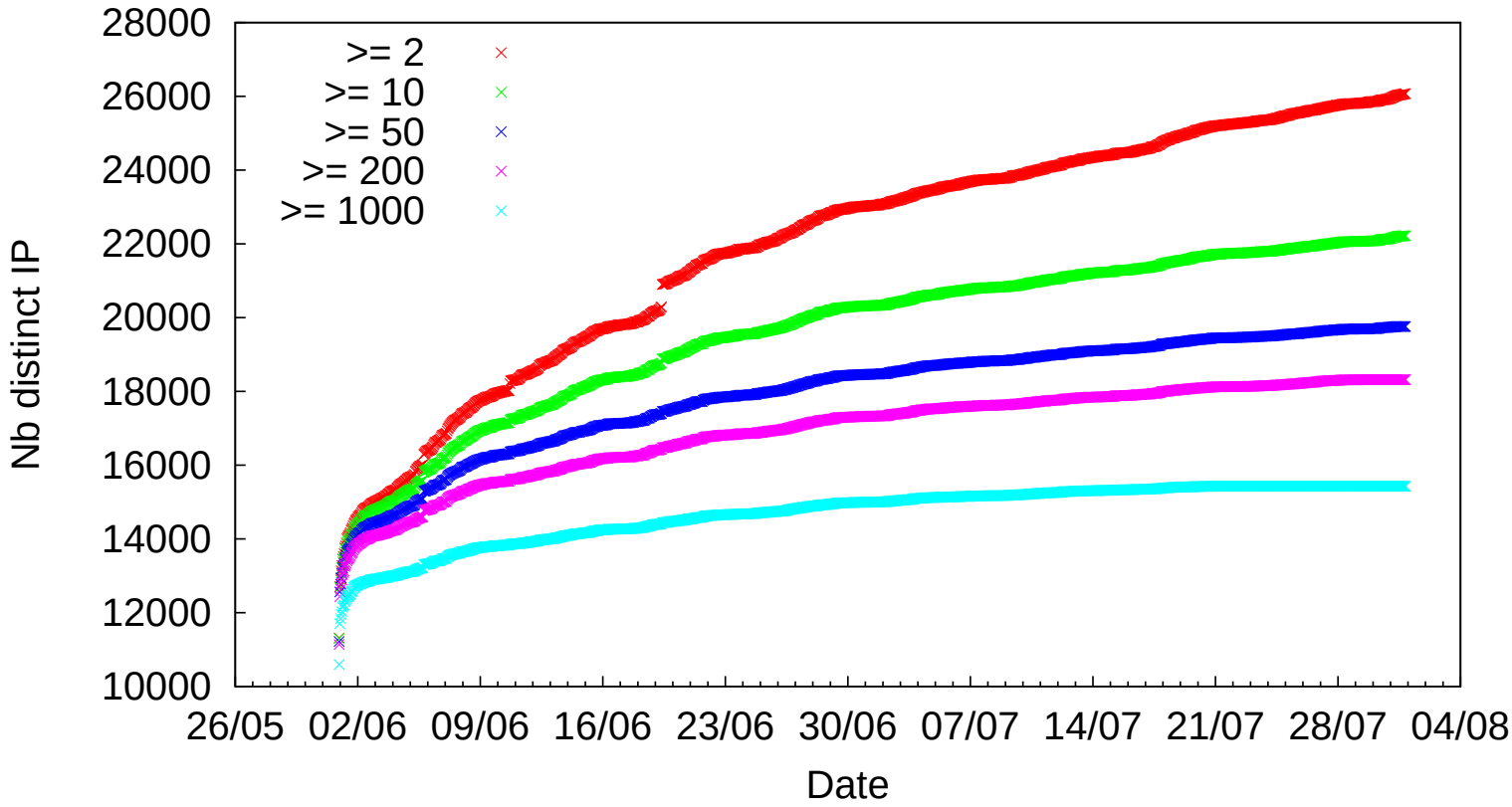
Influence des nœuds vus une seule fois ?



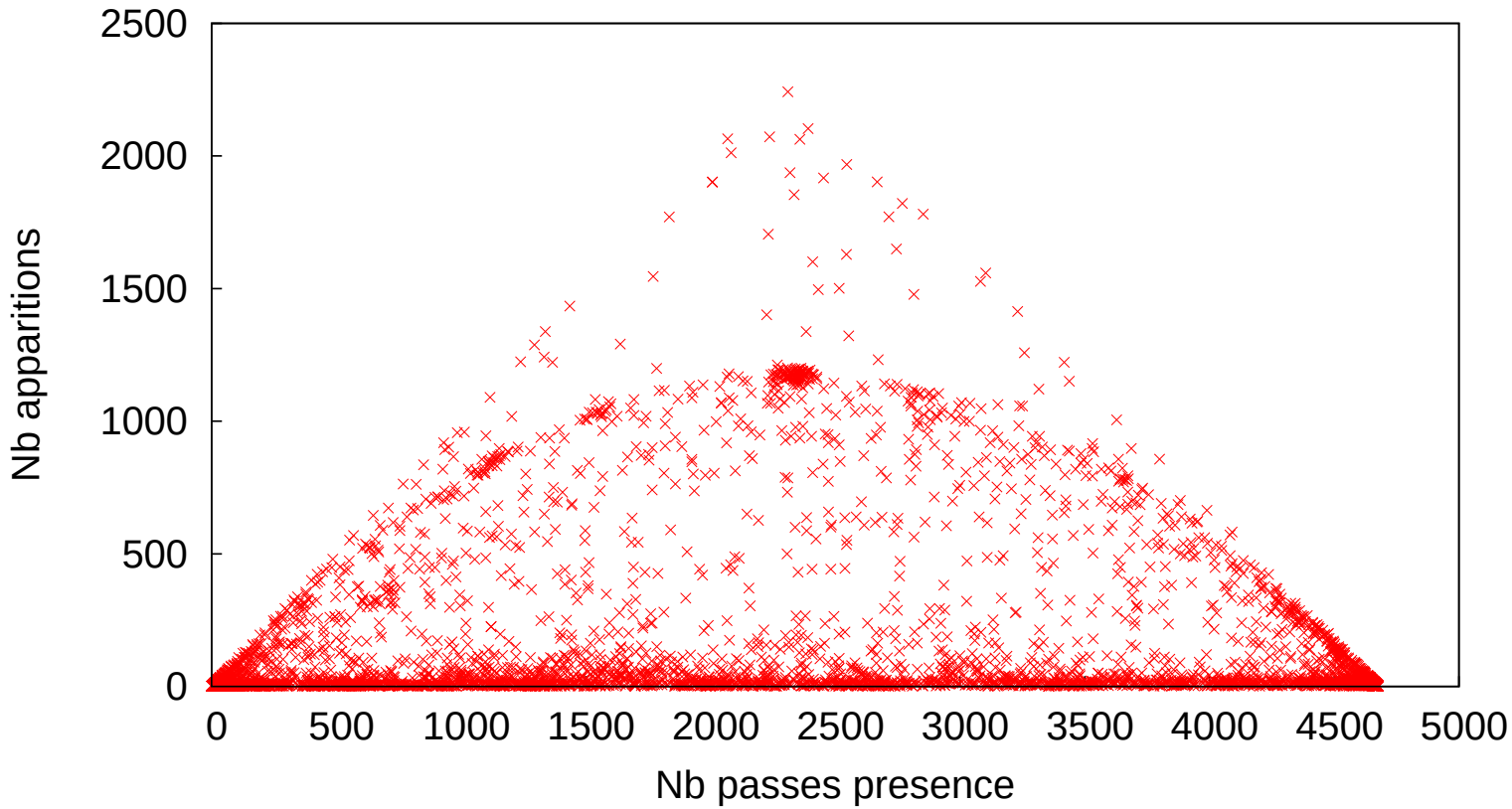
Influence des nœuds vus une seule fois ?



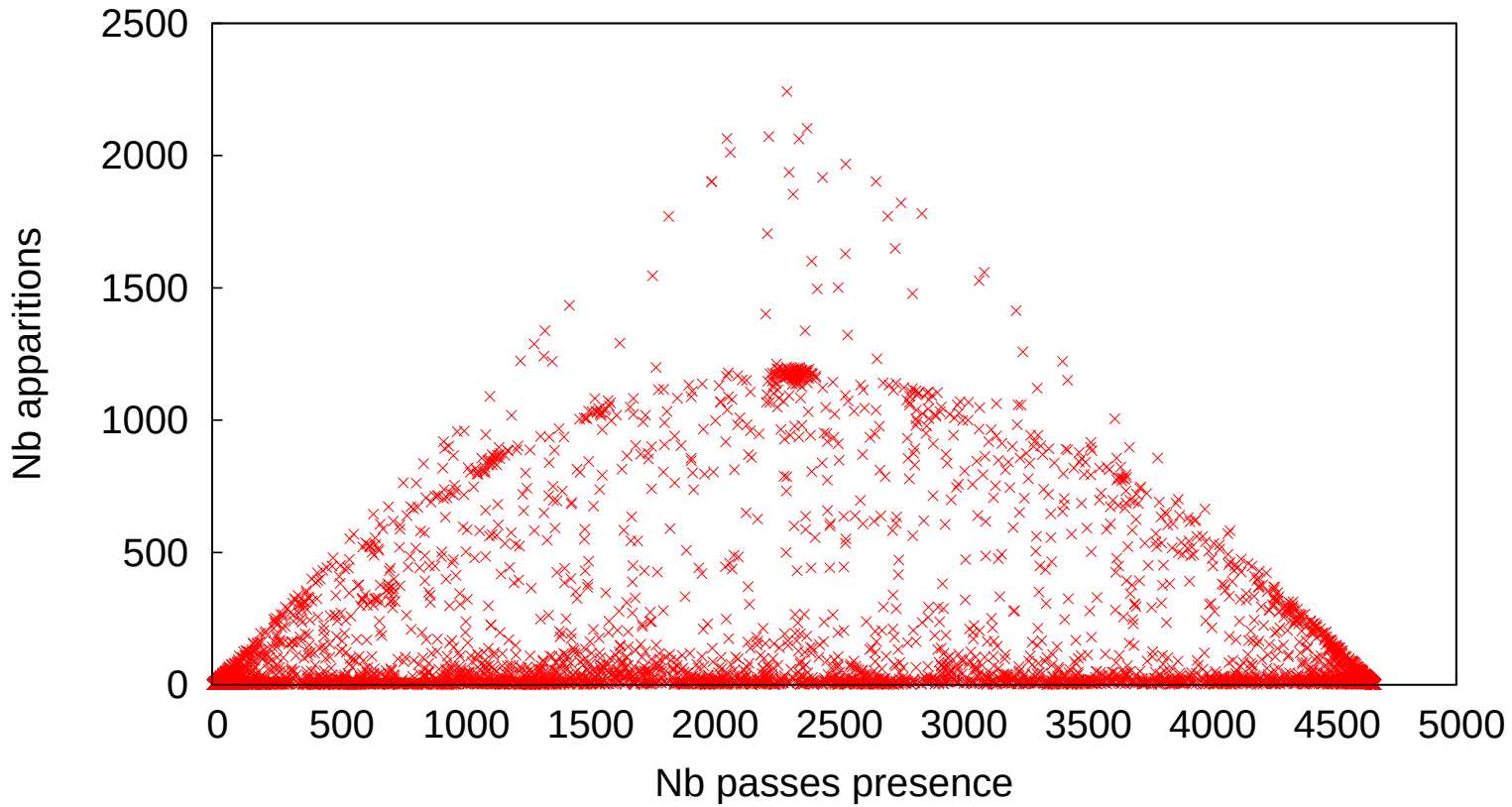
Influence des nœuds vus une seule fois ?



Présences vs apparitions

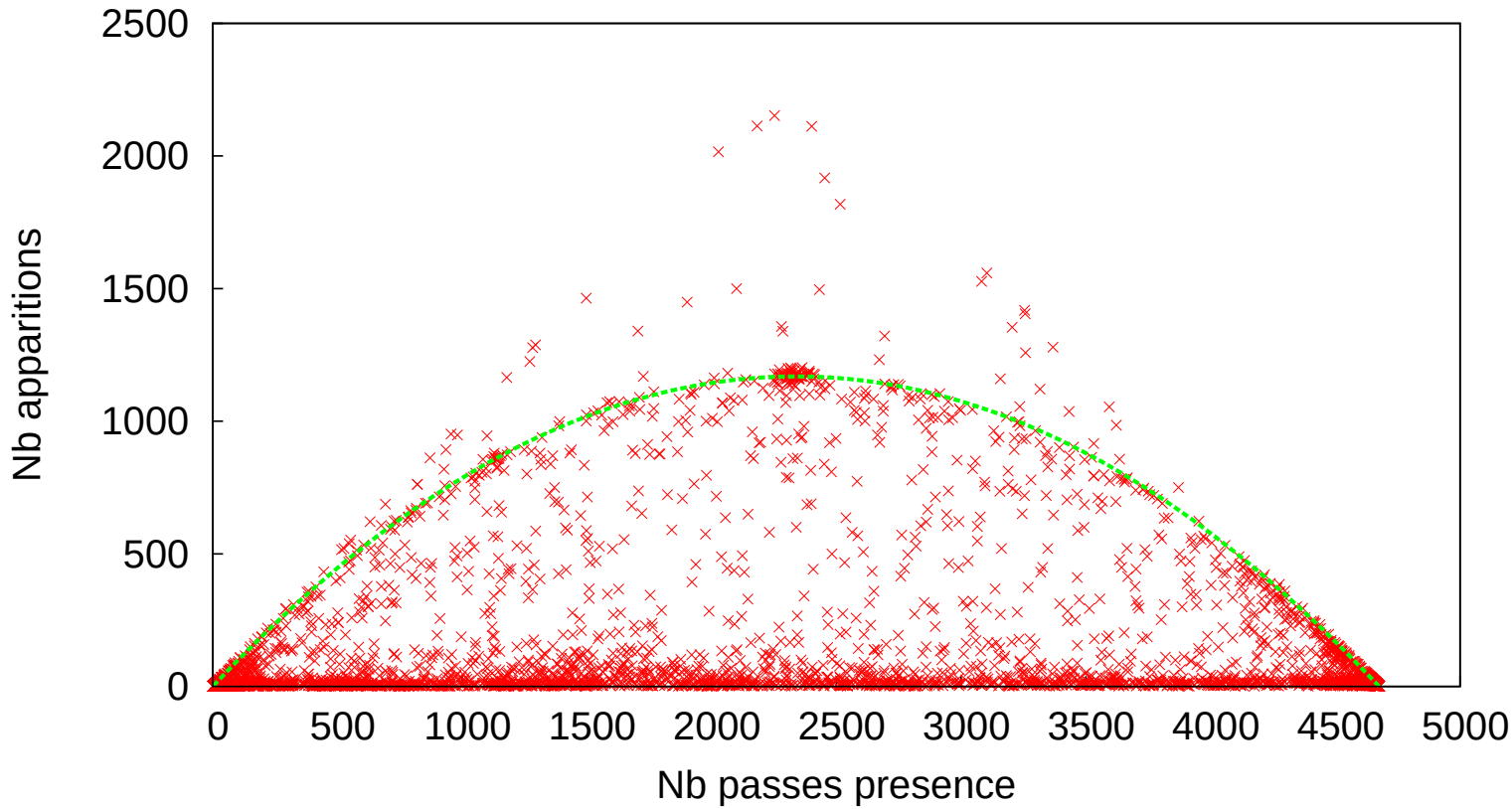


Présences vs apparitions



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

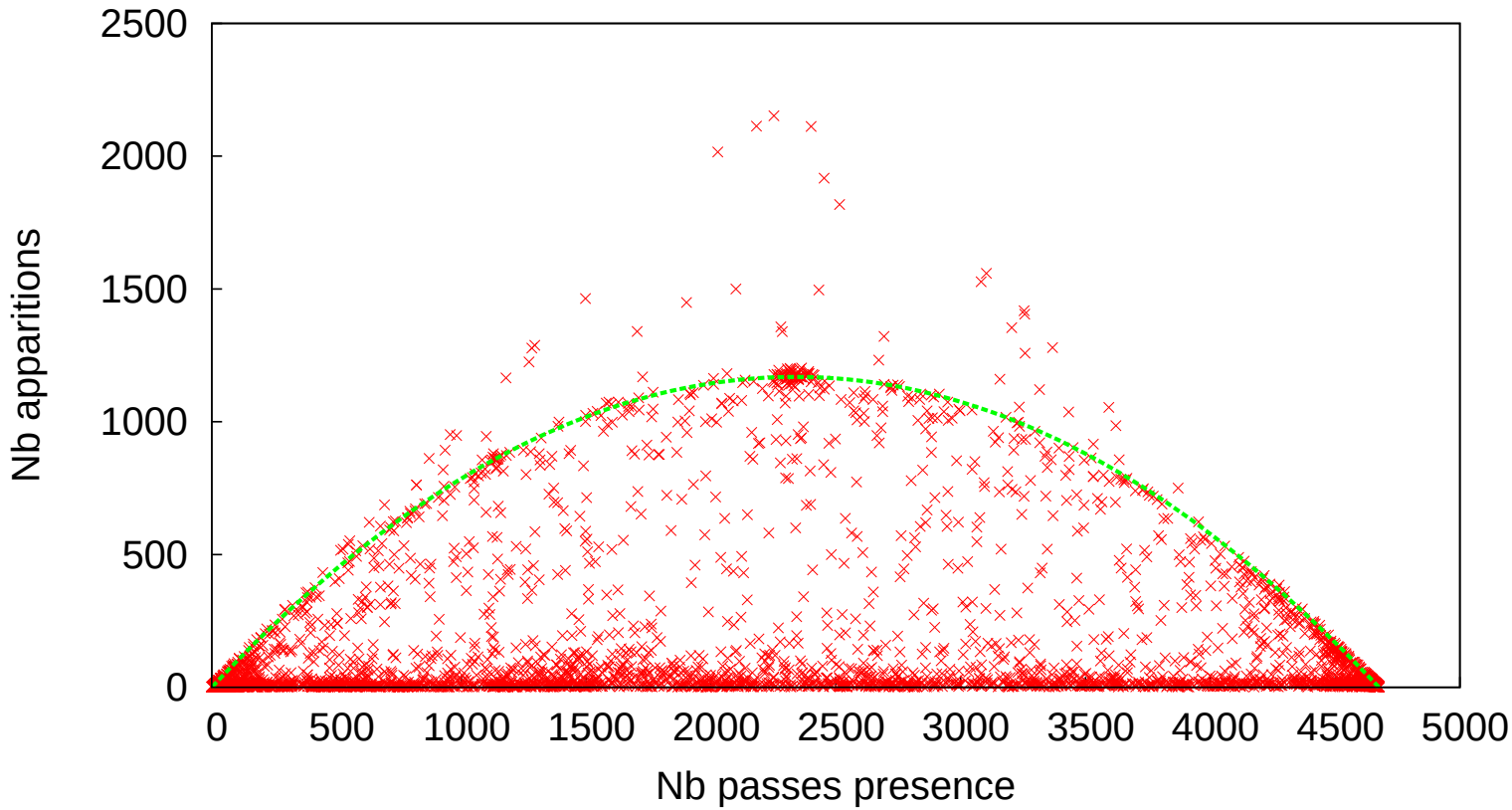
Présences vs apparitions



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

$$x(1-x)$$

Présences vs apparitions



Trois classes de nœuds

clignotent || aléatoires || majorité de stables

Évolution de l'ensemble des nœuds – Conclusion

- Intuition (renouvellement, présence, . . .)
- Notion de comportements des nœuds

Questions restantes

- Durée de vie, . . .
- Autres graphes $\rightarrow$ comportements généraux ?
- Comparaison à des modèles aléatoires simples ?

Attachement préférentiel

Idée :

les nouveaux liens s'attachent à des nœuds existants

$\hookrightarrow$ étudier les degrés de ces nœuds

Motivation : modèle d'Albert et Barábasi

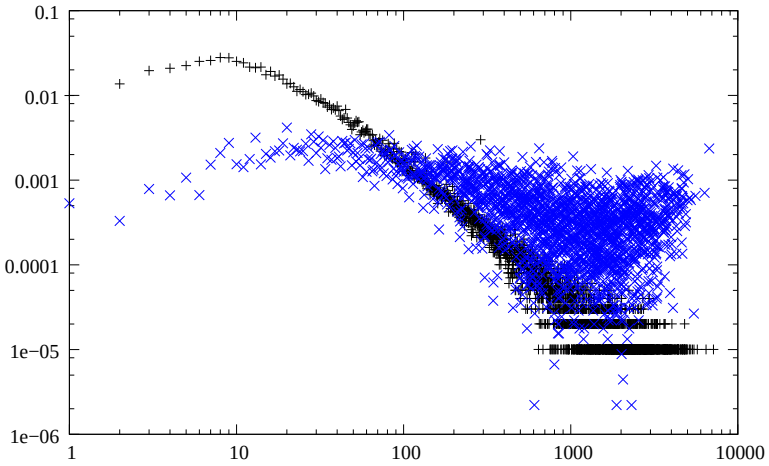
- un nouveau nœud par étape
- relié à un nœud existant
- choisi proportionnellement à son degré

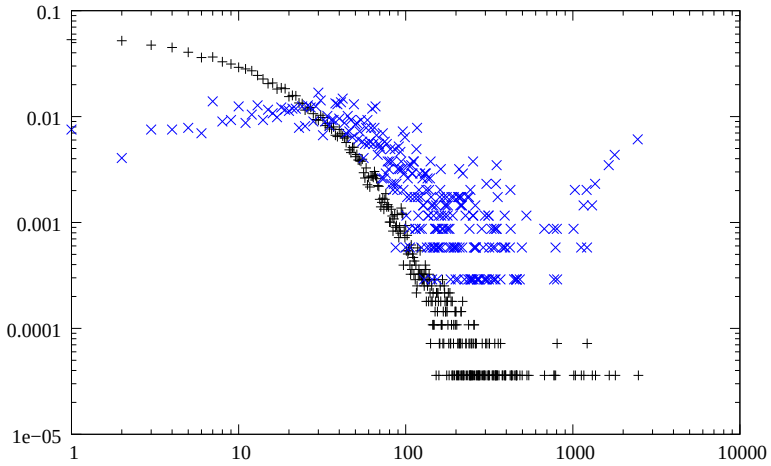
Idée : les nœuds de fort degré gagnent de plus en plus de liens.

Attachement préférentiel

Comparer :

- **Les degrés dans le graphe**
- **Les degrés des extrémités des nouveaux liens**





Conclusion sur l'attachement préférentiel

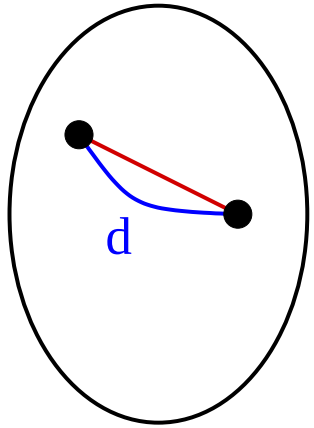
- Comportements différents de l'aléatoire
- Avec un biais vers les forts degrés

Formaliser ?

Clustering dans la dynamique

Idée :

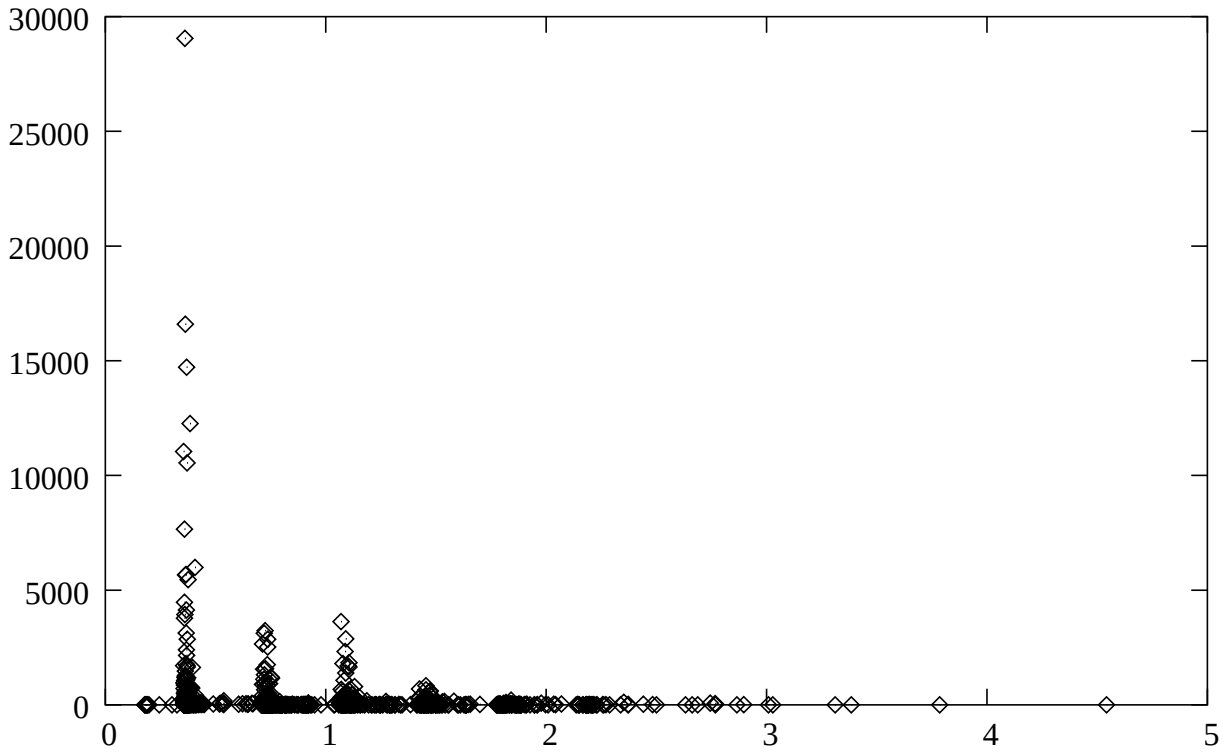
- les nouveaux liens relient des sommets
- regarder si ce sont des sommets proches



Clustering dans la dynamique

Interprétation ?

$$\frac{d}{d_{moy}}$$



Conclusion sur le clustering dans la dynamique

- les nouveaux liens relient des nœuds déjà proches
  $\hookrightarrow$ densité locale, clustering
- mais de temps en temps des nœuds éloignés
  $\rightarrow$ événement ?
  $\rightarrow$ moyen de les identifier formellement ?

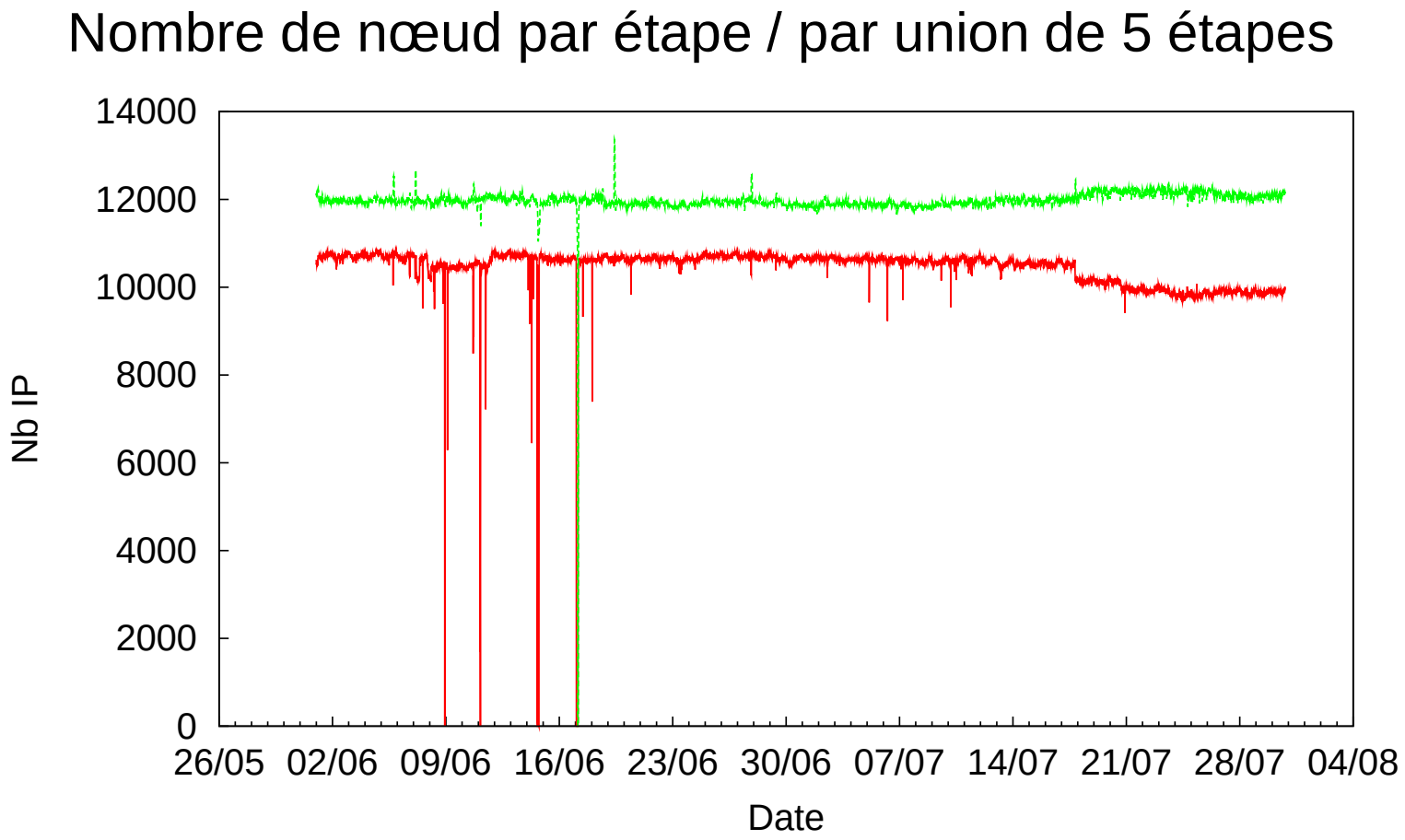
Conclusion sur le clustering dans la dynamique

- les nouveaux liens relient des nœuds déjà proches
  $\hookrightarrow$ densité locale, clustering
- mais de temps en temps des nœuds éloignés

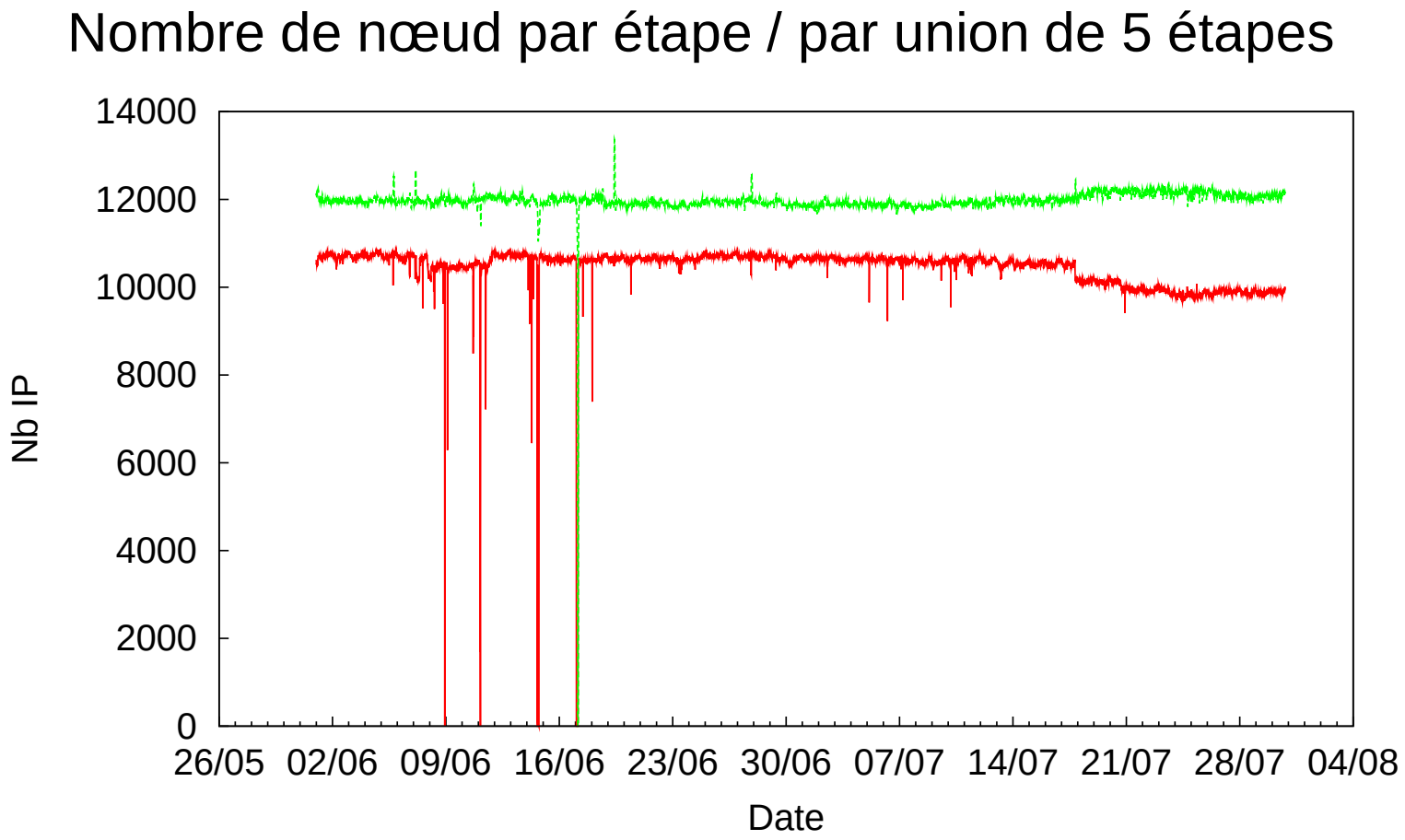
$\rightarrow$ événement ?

$\rightarrow$ moyen de les identifier formellement ?

Détection d'événements – Topologie de l'internet



Détection d'événements – Topologie de l'internet



pics → identification d'événements

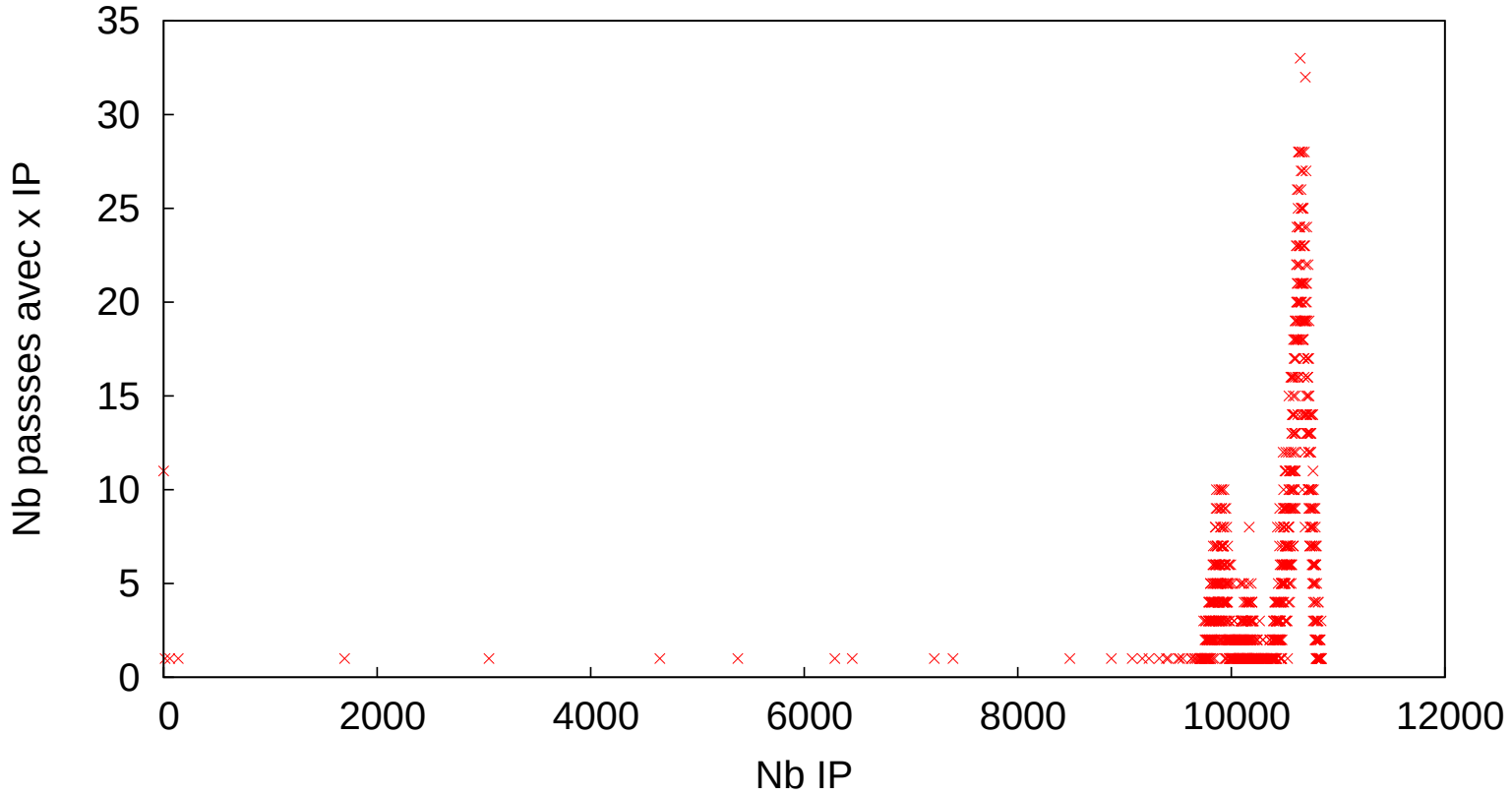
Identificiation formelle des pics ?

Pic $\longrightarrow$ valeur significativement plus grande que les autres.
Moyen de les identifier formellement ?

Identificiation formelle des pics ?

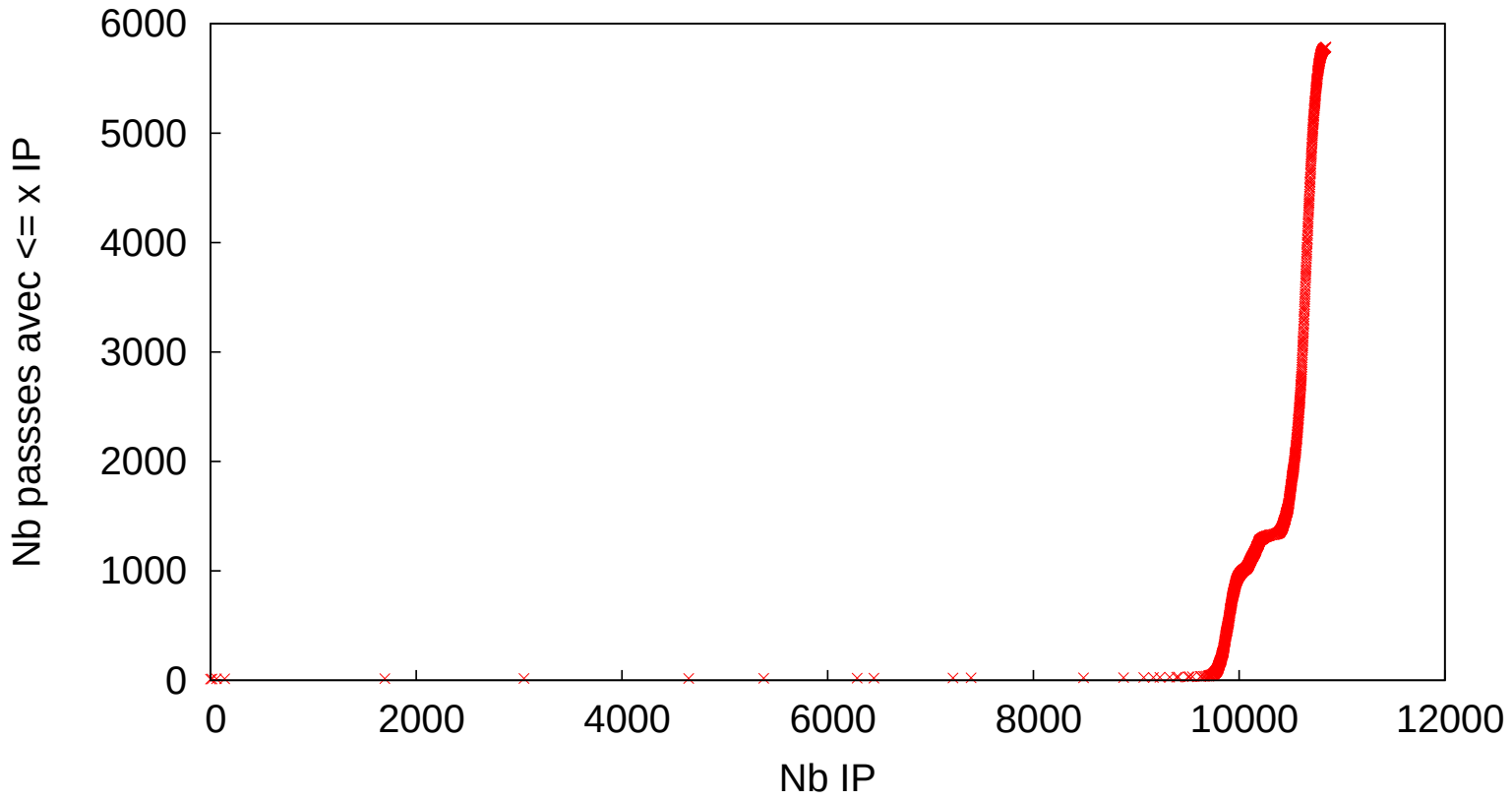
Distribution des valeurs observées

Nombre de nœud par étape



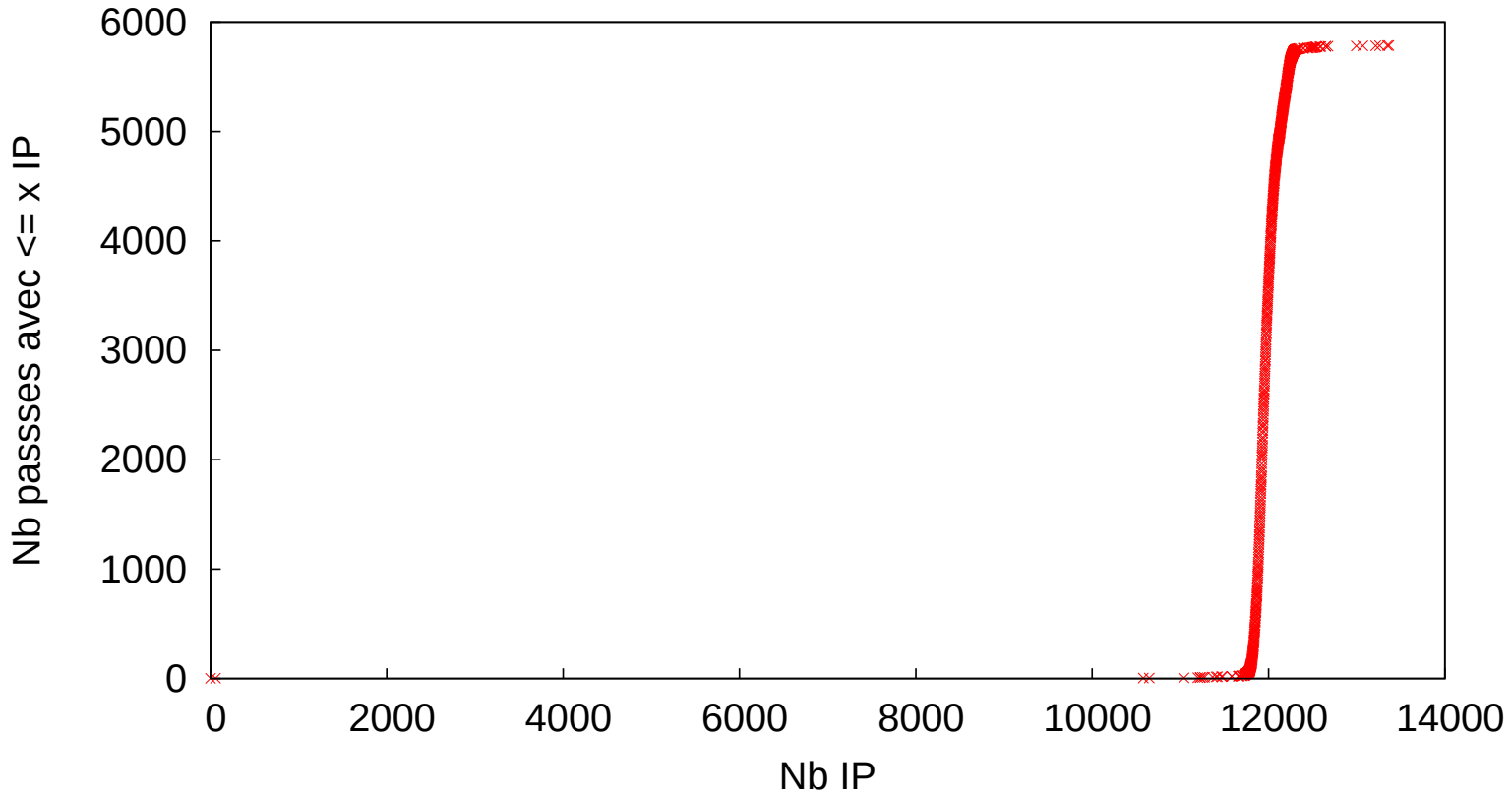
Identificiation formelle des pics ?

Distribution cumulative des valeurs observées
Nombre de nœud par étape



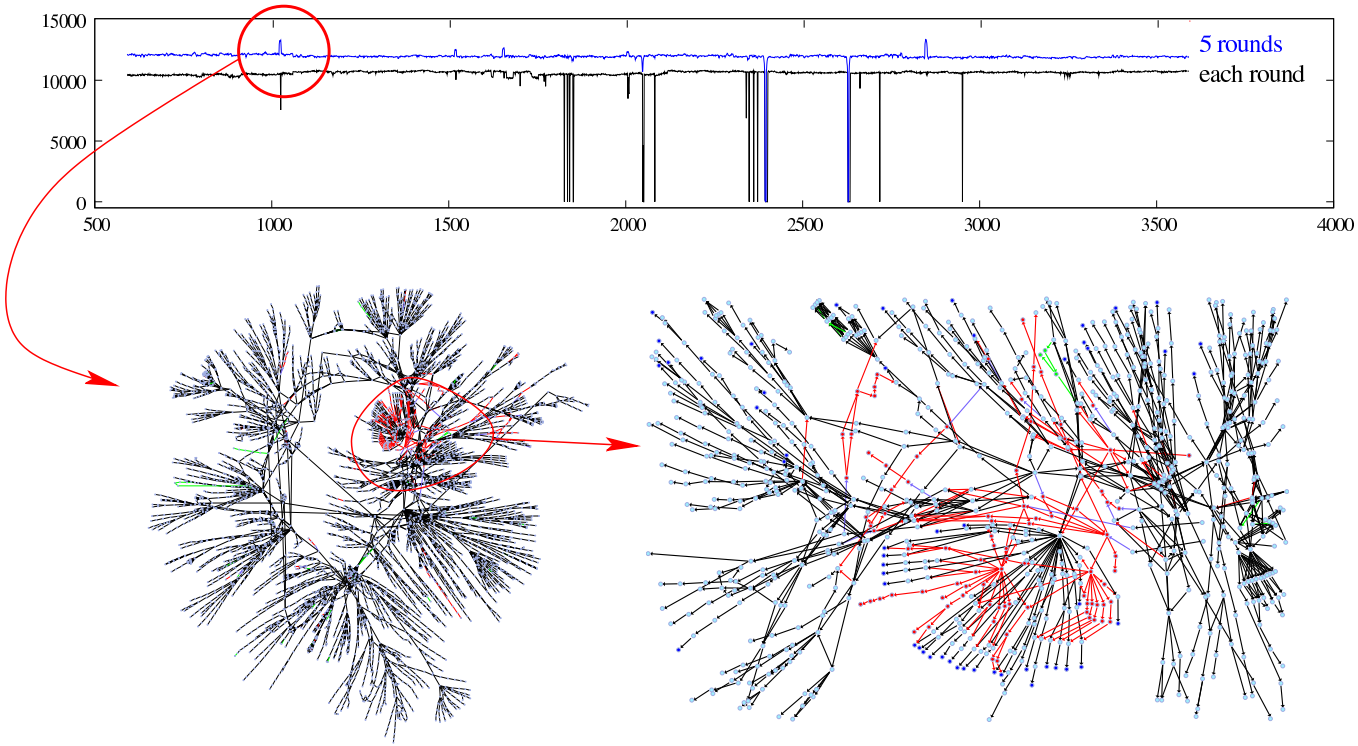
Identificiation formelle des pics ?

Distribution cumulative des valeurs observées
Nombre de nœuds dans 5 étapes consécutives



Valeurs hautes atypiques

Visualisation d'un événement



Événements – Résumé

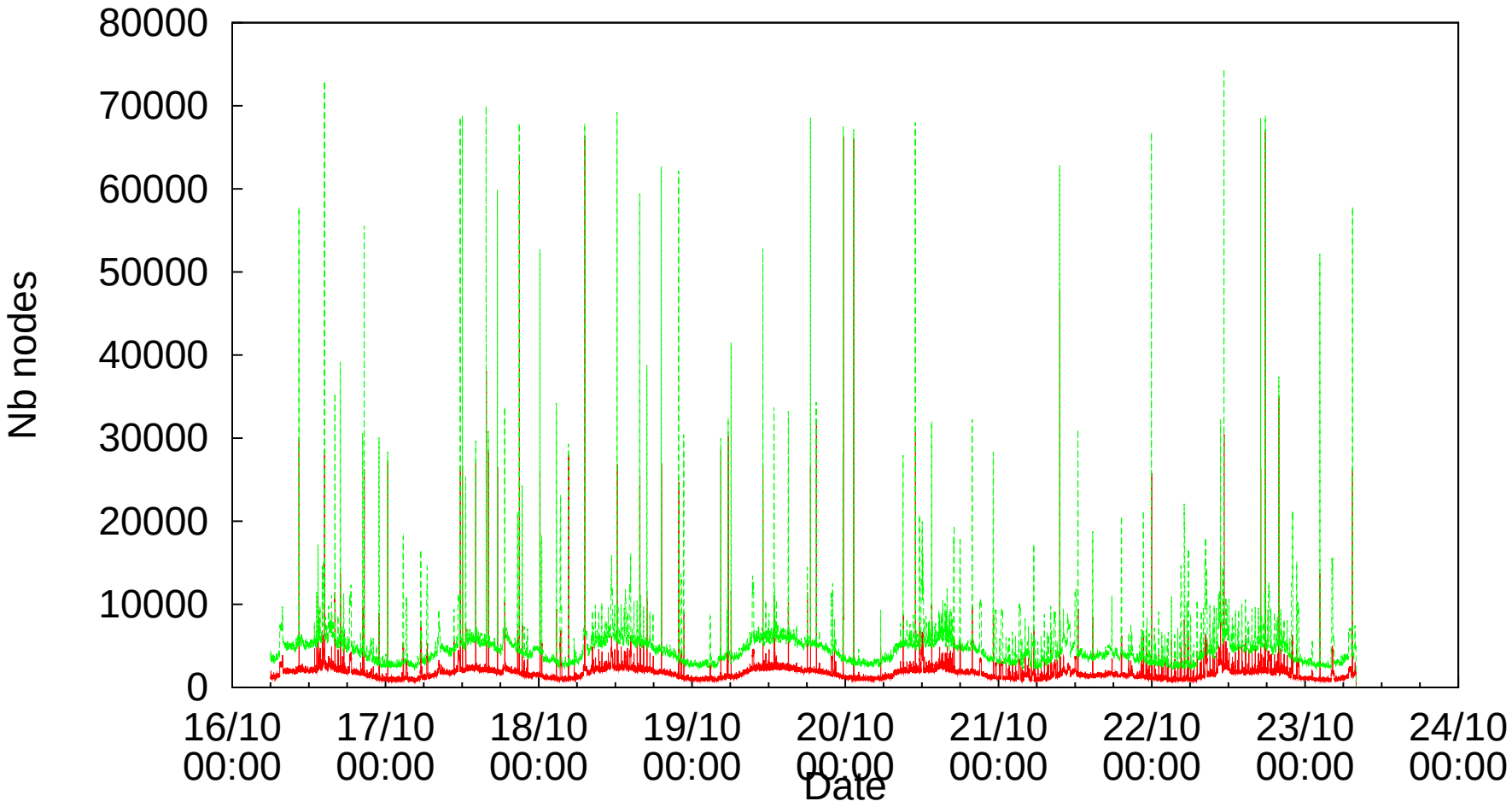
Technique pour détécter des événements.

Très dépendante

- Du graphe étudié
- De la façon de mesurer

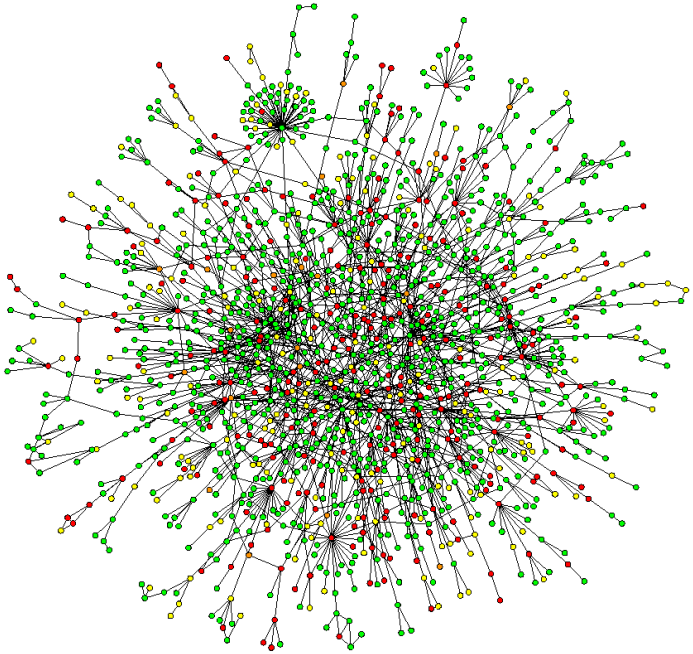
Pas adapté pour le trafic IP

Événements – Résumé



Visualisation

Exemple de graphe statique



Joli mais...

- On ne comprend rien
- Animer ?

Visualisation

Fichiers externes

Conclusion

Quelques pistes, nombreuses questions

- Échelle de temps
- Évolution de l'ensemble des nœuds
- Comportements des nœuds
- Détection d'événements
- Communautés
- Modélisation
- Questions spécifiques aux applications
- . . .

Conclusion

Équilibrer

- L'étude de questions spécifiques
- L'étude de graphes spécifiques

Collaborations pluridisciplinaires