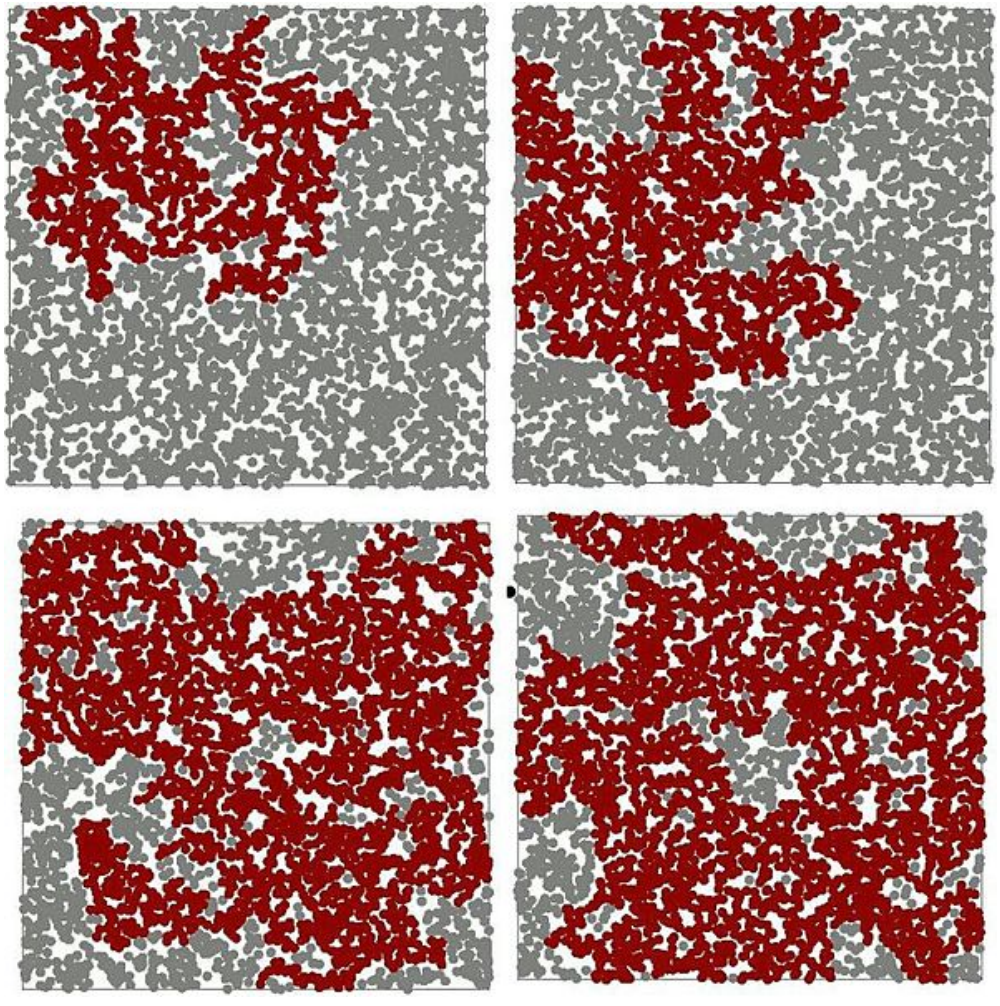
Les graphes aléatoires: un outil probabiliste pour l'informatique

Marc Lelarge

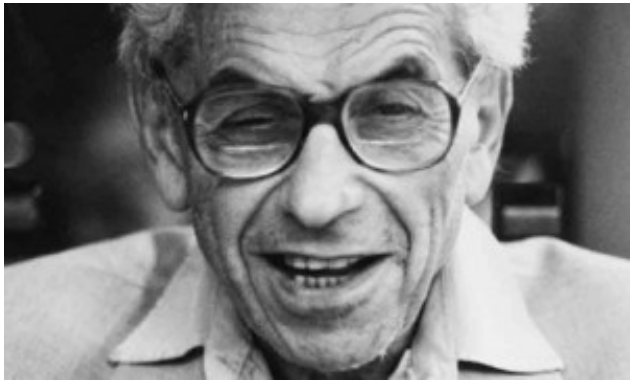
DYOGENE
INRIA-ENS

1/2 heure de Science, Dec. 5

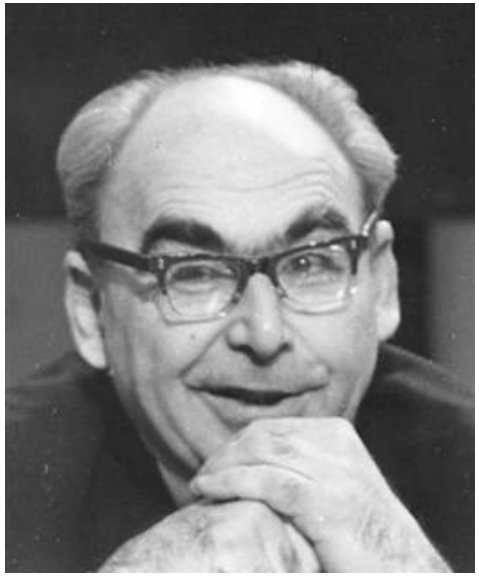
Ce dont je ne vais pas parler!



Graphe aléatoire d'Erdős, Rényi, Gilbert (1959)



Paul Erdős
(1913-1996)



Alfréd Rényi
(1921-1970)

Edgar Gilbert
(1923-2013)

Motivation: réseau téléphonique

RANDOM GRAPHS

By E. N. Gilbert

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

1. Introduction. Let N points, numbered 1, 2, $\cdots$, N, be given. There are $N(N-1)/2$ lines which can be drawn joining pairs of these points. Choosing a subset of these lines to draw, one obtains a graph; there are $2^{N(N-1)/2}$ possible graphs in total. Pick one of these graphs by the following random process. For all pairs of points make random choices, independent of each other, whether or not to join the points of the pair by a line. Let the common probability of joining be p. Equivalently, one may erase lines, with common probability $q = 1 - p$ from the complete graph.

In the random graph so constructed one says that *point i is connected to point j* if some of the lines of the graph form a path from i to j. If i is connected to j for every pair i, j, then the graph is said to be *connected*. The probability P_N that the graph is connected, and also the probability R_N that two specific points, say 1 and 2, are connected, will both be found.

As an application, imagine the N points to be N telephone central offices and suppose that each pair of offices has the same probability p that there is an idle direct line between them. Suppose further that a new call between two offices can be routed via other offices if necessary. Then R_N is the probability that there is some way of routing a new call from office 1 to office 2 and P_N is the probability that each office can call every other office.

Graphe aléatoire Erdős Rényi Gilbert

Commencer par le graphe vide sur n sommets et rajouter avec probabilité $p(n)$ chacune des $\binom{n}{2}$ arêtes de manière indépendante.
On obtien ainsi $G(n,p)$.

- Nombre moyen d'arêtes: $p\frac{n(n-1)}{2}$ donc degré moyen: $\sim pn$.
- Si $p > (1+\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est connecté a.g.p. et si $p < (1-\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est déconnecté a.g.p.
- Si $np > 1$ alors $G(n,p)$ contient une composante géante de taille linéaire en n; si $np = 1$ alors $G(n,p)$ contient une composante de taille $\Theta(n^{2/3})$; si $np < 1$ alors la plus grande composante de $G(n,p)$ est d'ordre $\ln n$.

Graphe aléatoire Erdős Rényi Gilbert

Commencer par le graphe vide sur n sommets et rajouter avec probabilité $p(n)$ chacune des $\binom{n}{2}$ arêtes de manière indépendante.
On obtien ainsi $G(n,p)$.

- Nombre moyen d'arêtes: $p\frac{n(n-1)}{2}$ donc degré moyen: $\sim pn$.
- Si $p > (1+\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est connecté a.g.p. et si $p < (1-\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est déconnecté a.g.p.
- Si $np > 1$ alors $G(n,p)$ contient une composante géante de taille linéaire en n; si $np = 1$ alors $G(n,p)$ contient une composante de taille $\Theta(n^{2/3})$; si $np < 1$ alors la plus grande composante de $G(n,p)$ est d'ordre $\ln n$.

Graphe aléatoire Erdős Rényi Gilbert

Commencer par le graphe vide sur n sommets et rajouter avec probabilité $p(n)$ chacune des $\binom{n}{2}$ arêtes de manière indépendante.
On obtien ainsi $G(n,p)$.

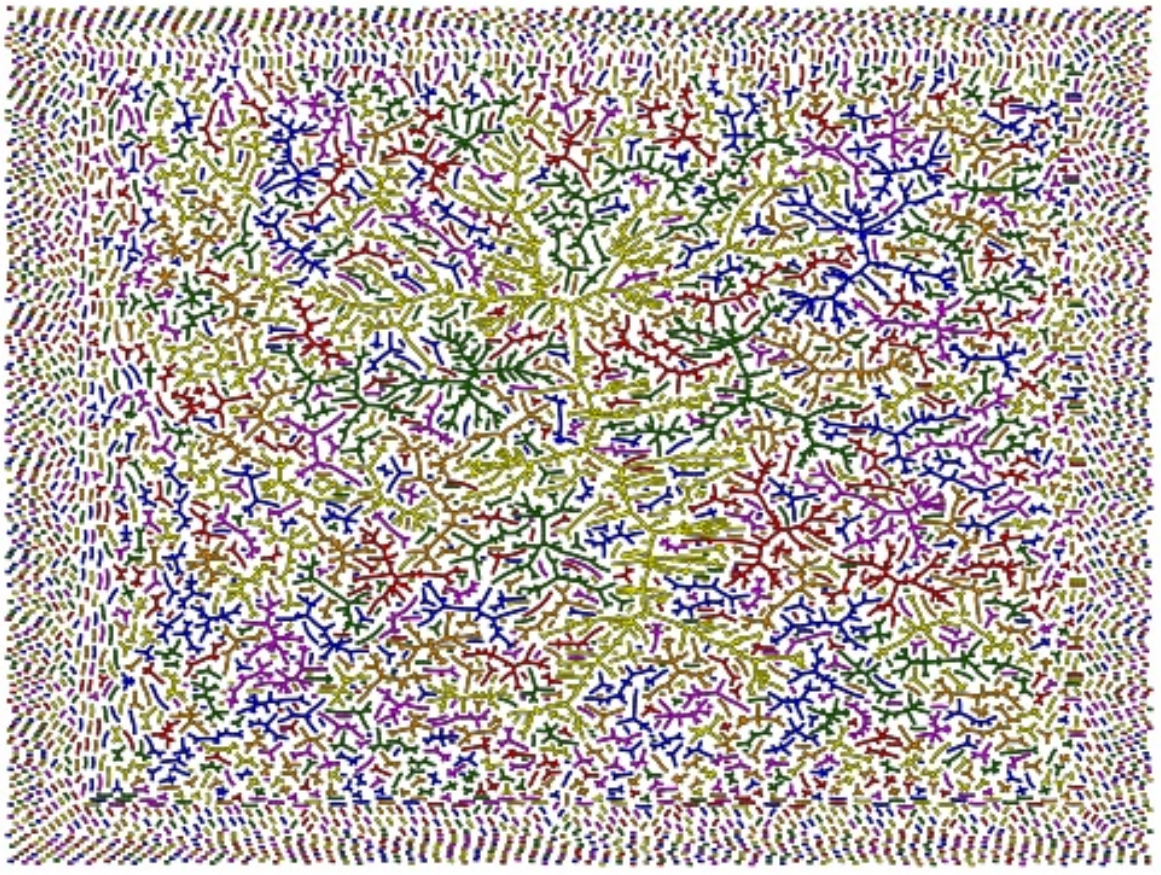
- Nombre moyen d'arêtes: $p\frac{n(n-1)}{2}$ donc degré moyen: $\sim pn$.
- Si $p > (1+\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est connecté a.g.p. et si $p < (1-\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est déconnecté a.g.p.
- Si $np > 1$ alors $G(n,p)$ contient une composante géante de taille linéaire en n; si $np = 1$ alors $G(n,p)$ contient une composante de taille $\Theta(n^{2/3})$; si $np < 1$ alors la plus grande composante de $G(n,p)$ est d'ordre $\ln n$.

Graphe aléatoire Erdős Rényi Gilbert

Commencer par le graphe vide sur n sommets et rajouter avec probabilité $p(n)$ chacune des $\binom{n}{2}$ arêtes de manière indépendante.
On obtien ainsi $G(n,p)$.

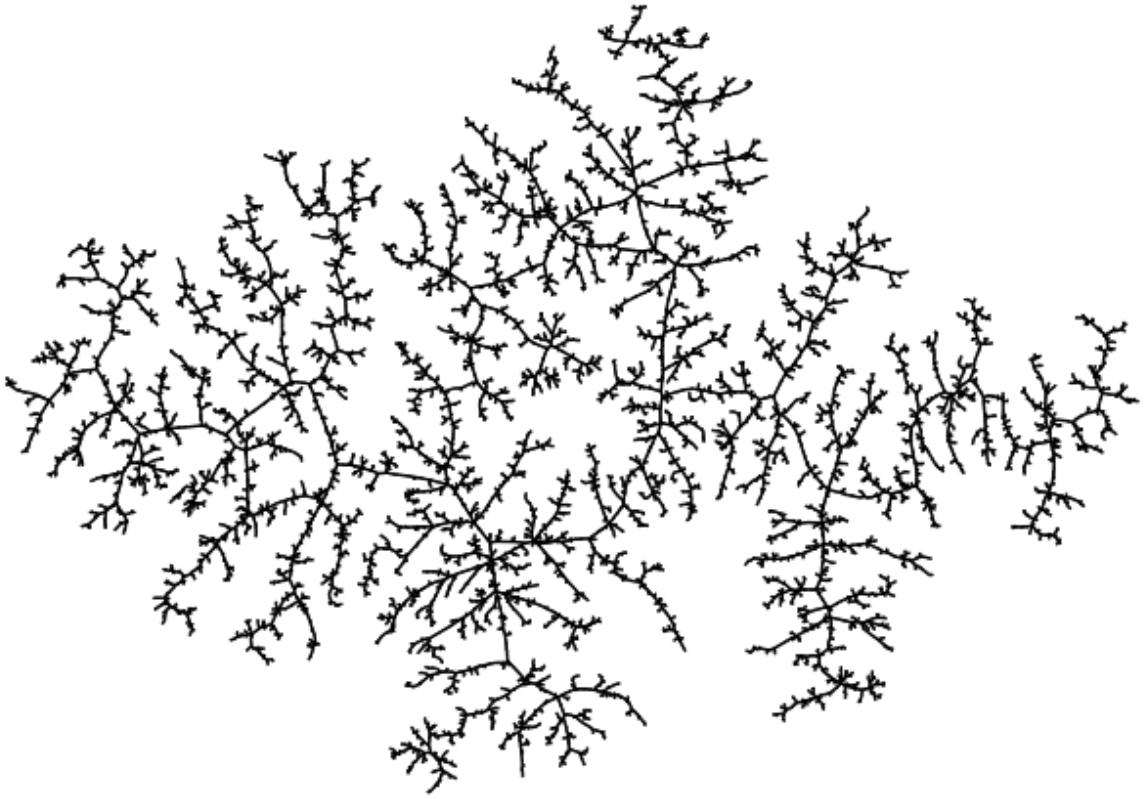
- Nombre moyen d'arêtes: $p\frac{n(n-1)}{2}$ donc degré moyen: $\sim pn$.
- Si $p > (1+\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est connecté a.g.p. et si $p < (1-\epsilon)\frac{\ln n}{n}$ alors $G(n,p)$ est déconnecté a.g.p.
- Si $np > 1$ alors $G(n,p)$ contient une composante géante de taille linéaire en n; si $np = 1$ alors $G(n,p)$ contient une composante de taille $\Theta(n^{2/3})$; si $np < 1$ alors la plus grande composante de $G(n,p)$ est d'ordre $\ln n$.

Autour de la transition de phase $p \sim \frac{1}{n}$

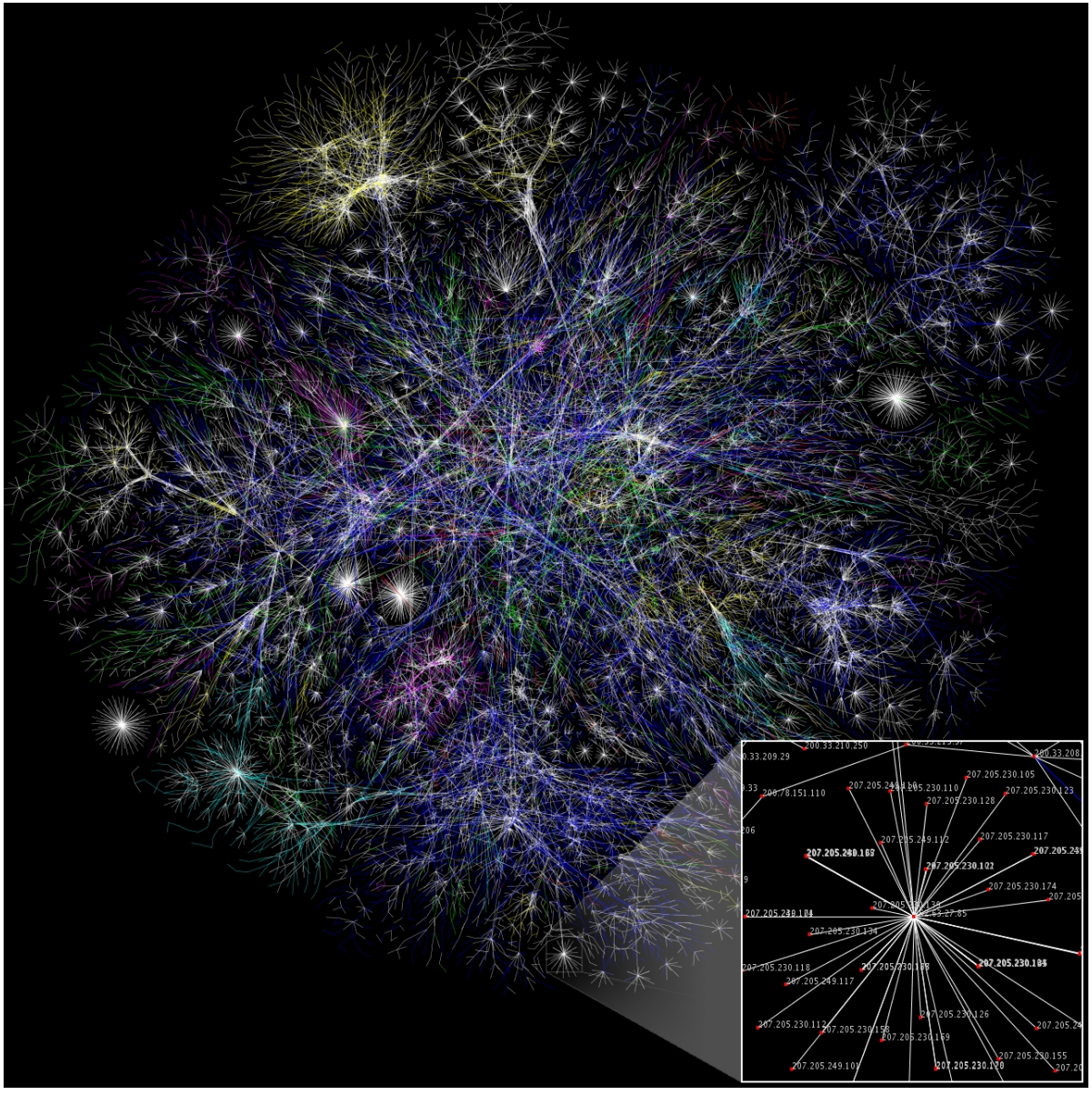


Autour de la transition de phase $p \sim \frac{1}{n}$

La composante critique:



’Complex Network’



Détection de communautés

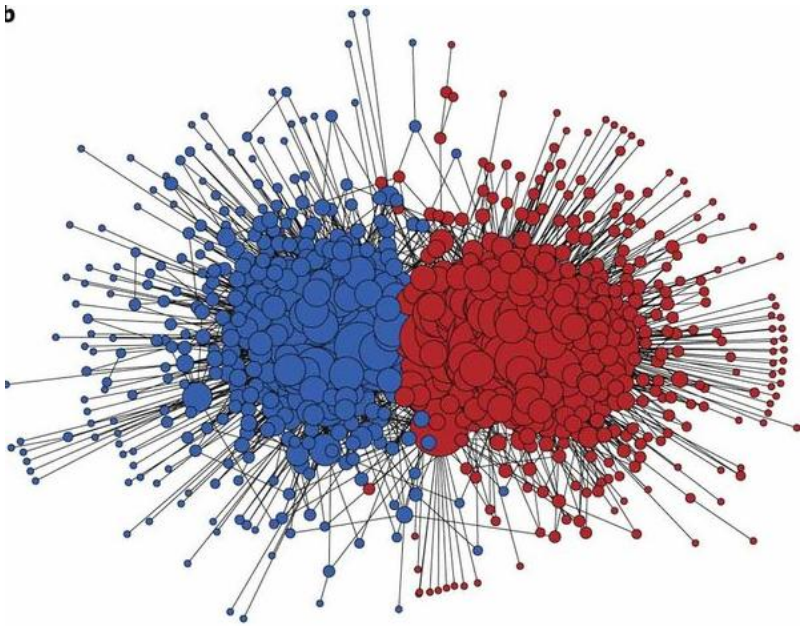
- Un graphe aléatoire est un modèle sans structure qui sert de graphe 'témoin'.
- La modularité est la différence entre la densité d'arêtes entre communautés et la densité moyenne correspondante dans un graphe aléatoire.
- Notion introduite par M.E.J. Newman (2004)

Détection de communautés

- Un graphe aléatoire est un modèle sans structure qui sert de graphe 'témoin'.
- La modularité est la différence entre la densité d'arêtes entre communautés et la densité moyenne correspondante dans un graphe aléatoire.
- Notion introduite par M.E.J. Newman (2004)

Détection de communautés

- Un graphe aléatoire est un modèle sans structure qui sert de graphe 'témoin'.
- La modularité est la différence entre la densité d'arêtes entre communautés et la densité moyenne correspondante dans un graphe aléatoire.
- Notion introduite par M.E.J. Newman (2004)



Méthode spectrale

- Idée: cacher des communautés dans un graphe aléatoire.
- But: montrer qu'un algorithme spectral retrouve les communautés.
- Technique de preuve: la matrice d'adjacence peut être vue comme une version bruitée d'un signal de faible dimmension associé aux communautés.

Méthode spectrale

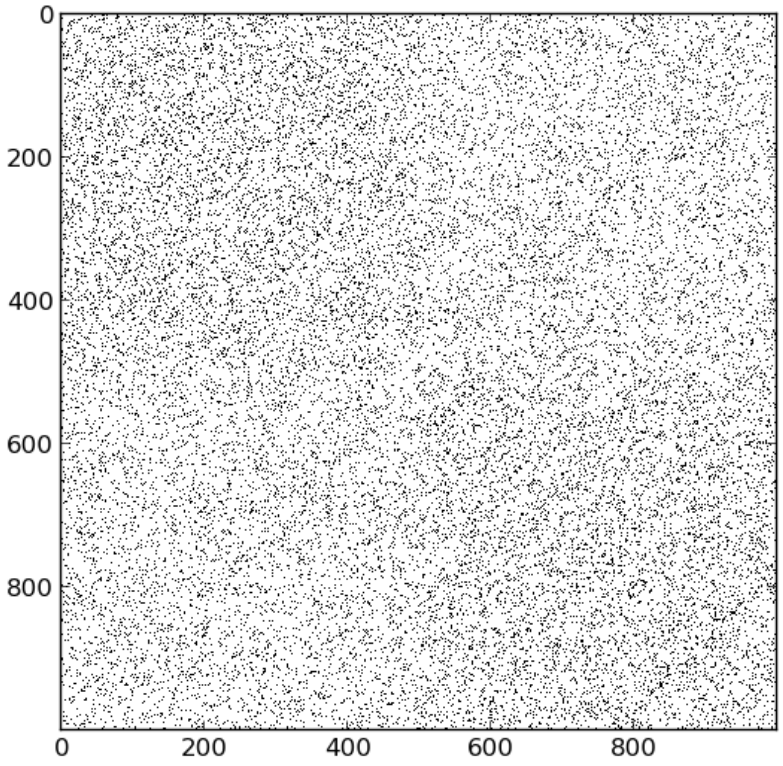
- Idée: cacher des communautés dans un graphe aléatoire.
- But: montrer qu'un algorithme spectral retrouve les communautés.
- Technique de preuve: la matrice d'adjacence peut être vue comme une version bruitée d'un signal de faible dimmension associé aux communautés.

Méthode spectrale

- Idée: cacher des communautés dans un graphe aléatoire.
- But: montrer qu'un algorithme spectral retrouve les communautés.
- Technique de preuve: la matrice d'adjacence peut être vue comme une version bruitée d'un signal de faible dimmension associé aux communautés.

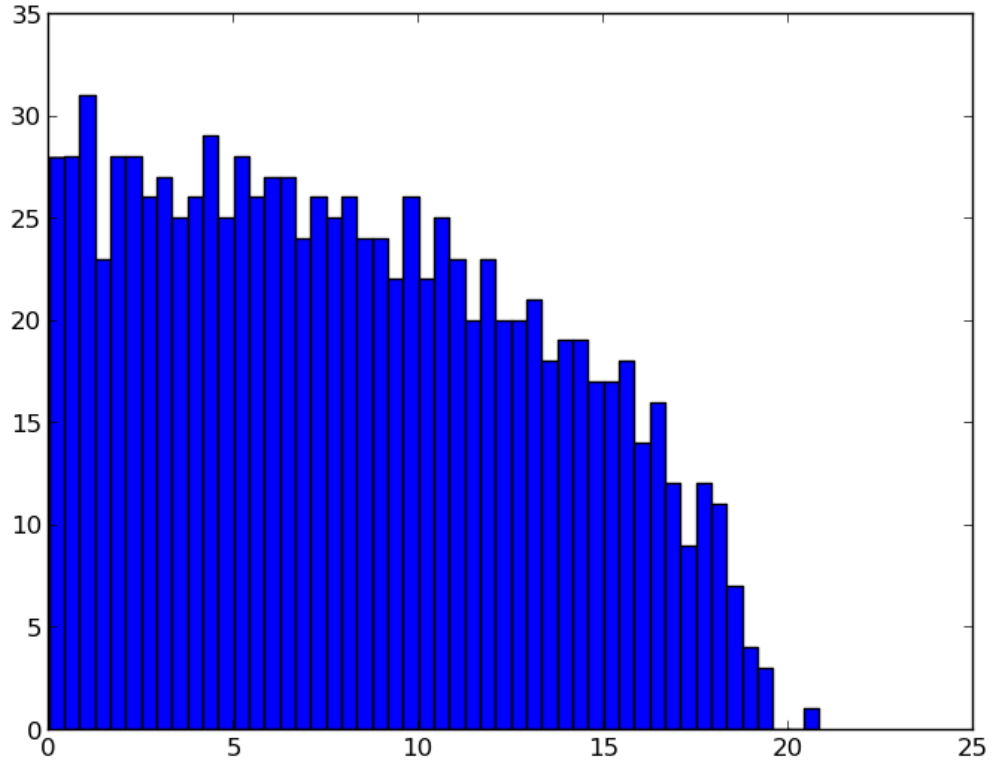
Méthode spectrale en pratique

Matrice d'adjacence avec 2 communautés.



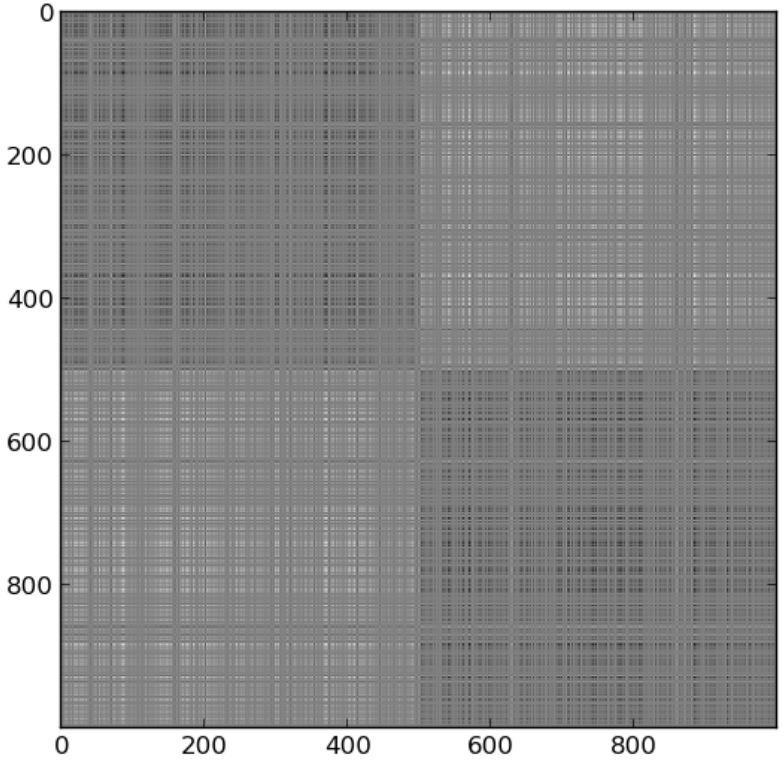
Méthode spectrale en pratique

Spèctre de la matrice d'adjacence avec 2 communautés.



Méthode spectrale en pratique

Approximation de rang 1 de la matrice d'adjacence centrée.



Méthode spectrale: résultats

Proposition

L'algorithme spectral est optimal pour la famille des graphes aléatoires ayant des communautés 'plantées' dès que le degré moyen est d'ordre logarithmique en n.

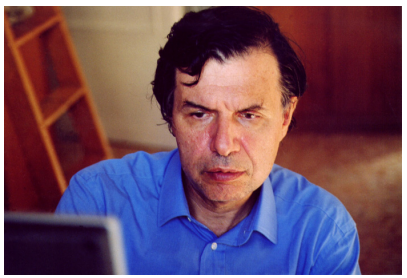
Travail en commun avec L. Massoulié (MSR-INRIA) et J. Xu (UIUC)

Modèle dense: Sherrington-Kirkpatrick (1975)

- Modèle de verre de spin sur le graphe complet. Hamiltonien donné par ($J_{ij} \sim \mathcal{N}(0,1)$):

$$H = \frac{1}{\sqrt{n}} \sum_{i<j} J_{ij}\sigma_i\sigma_j$$

- Résolu par Giorgio Parisi (1979)



- Puis preuve rigoureuse par Michel Talagrand (2006)

Méthode de la cavité

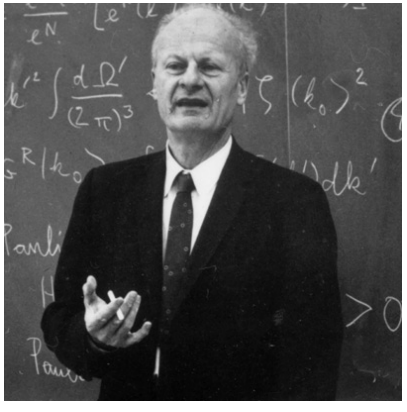
- Développée en physique statistique pour l'étude du modèle SK.
- Idée similaire à la programmation dynamique: en créant une cavité, interpréter l'influence du reste du graphe comme des 'messages indépendants' puis itérer.
- Adaptée au cas des graphes dilués: Bethe lattice (1935).

Méthode de la cavité

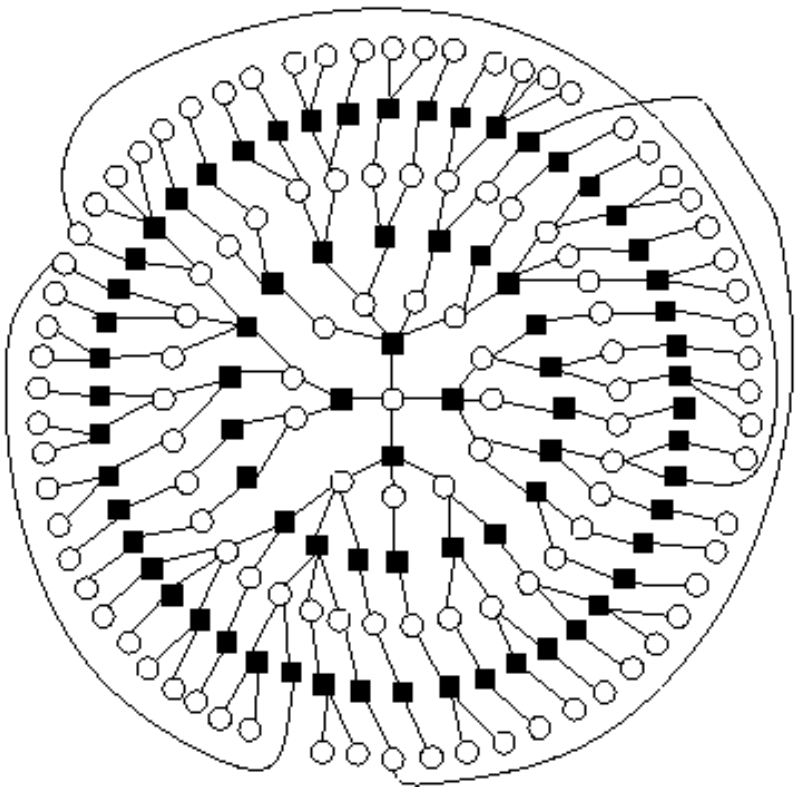
- Développée en physique statistique pour l'étude du modèle SK.
- Idée similaire à la programmation dynamique: en créant une cavité, interpréter l'influence du reste du graphe comme des 'messages indépendants' puis itérer.
- Adaptée au cas des graphes dilués: Bethe lattice (1935).

Méthode de la cavité

- Développée en physique statistique pour l'étude du modèle SK.
- Idée similaire à la programmation dynamique: en créant une cavité, interpréter l'influence du reste du graphe comme des 'messages indépendants' puis itérer.
- Adaptée au cas des graphes dilués: Bethe lattice (1935).



La méthode de la cavité pour les LDPC



Robert Gallager (1963) T. Richardson et R . Urbanke (2007)

Une autre application: équilibrage de charge et fonction de hachage

- m balles et n boîtes
- chaque balle choisit une boîte uniformément au hasard
- But: éviter les collisions.

C'est le problème des anniversaires. La probabilité qu'il n'y ait pas de collision est:

$$\begin{aligned} p(n,m) &= \left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-m+1}{n}\right) \\ &\approx \exp\left(-\frac{1+2+\cdots+m-1}{n}\right) \\ &\approx \exp\left(-\frac{m^2}{2n}\right) \end{aligned}$$

Donc pour éviter les collisions, il faut

$$p(n,m) \approx 1 \quad \Leftrightarrow \quad m << \sqrt{n}.$$

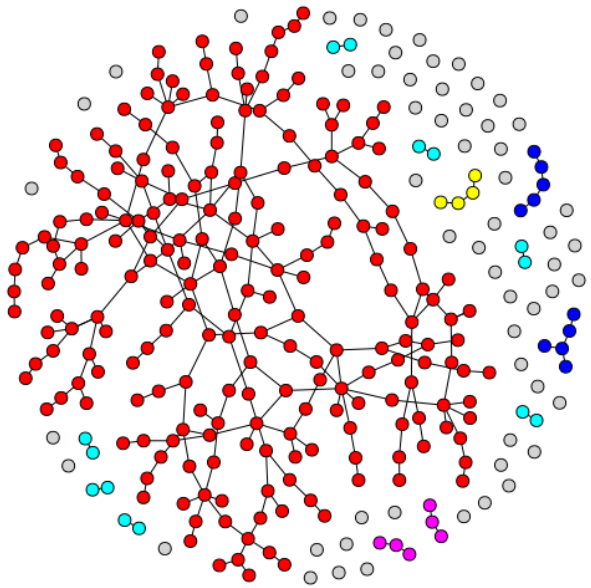
Cuckoo hashing

Introduit par Pagh et Rodler (2001):

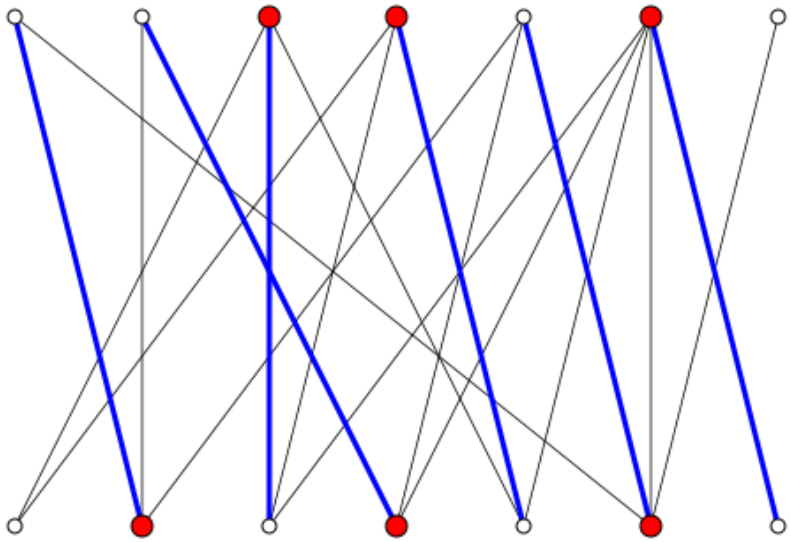
- 2 boîtes sont associées de manière aléatoire à chaque balle
- chaque balle doit être placée dans une de ces deux boîtes
- chaque boîte a capacité 1, i.e. aucune collision autorisée.

The power of two choices

En interprétant chaque boîte comme un sommet et les deux choix donnés à une balle comme une arête, on obtient le graphe aléatoire Erdős Rényi Gilbert!
La capacité du système est une question d'orientation du graphe. Dès qu'il y a une composante géante, il y a des collisions donc on doit avoir $m < \frac{n}{2}$ à comparer à $\sqrt{n}$ pour un seul choix!



Une généralisation: plus de choix



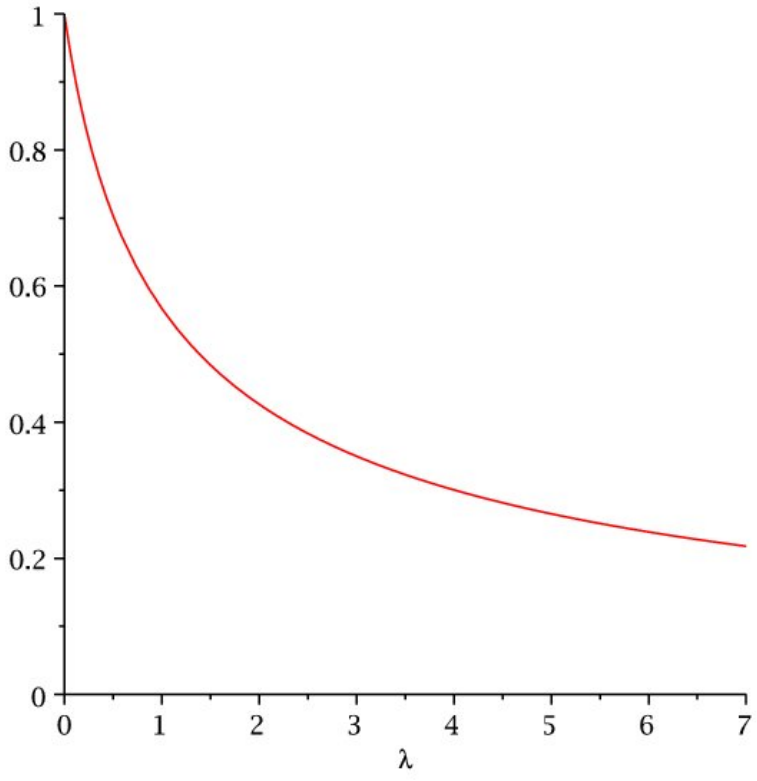
La capacité du système est une question de couplage dans le graphe biparti aléatoire.

La méthode de la cavité pour le couplage

Considérer un arbre aléatoire: processus de branchement où chaque individu à un nombre d'enfants aléatoire suivant une loi de Poisson de moyenne λ.
Algorithme glouton: si un enfant est disponible, mettre l'arête correspondante dans le couplage.
Calcul de la probabilité p que la racine ne soit couplée avec aucun de ses enfants, i.e. que tous ses enfants soient déjà couplés:

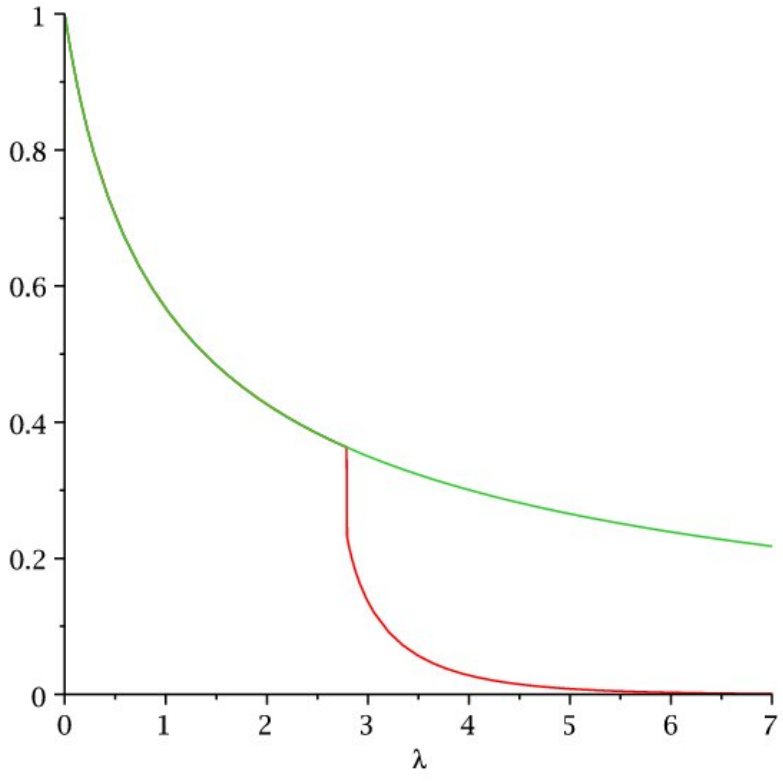
$$p \quad = \quad \mathbb{P}\,(\text{ aucun enfant disponible }) = e^{-\lambda p}$$

Une estimation naïve



Solution de l'équation $p = e^{-\lambda p}$ en fonction de λ.

Vraie valeur



Fonction $\lambda \mapsto p(\lambda)$.

Explications

Soit p_k la probabilité que la racine ne soit couplée avec aucun de ses enfants dans un arbre tronqué à profondeur k.

- $p_0 = 1$
- $p_1 = e^{\lambda}$
- puis pour $k \geq 0$

$$p_{k+1} = e^{-\lambda p_k}.$$

Nous avons calculé le point fixe de $p \mapsto e^{-\lambda p}$ mais en fait nous itérons cette fonction.

Explications

Soit p_k la probabilité que la racine ne soit couplée avec aucun de ses enfants dans un arbre tronqué à profondeur k.

- $p_0 = 1$
- $p_1 = e^{\lambda}$
- puis pour $k \geq 0$

$$p_{k+1} = e^{-\lambda p_k}.$$

Nous avons calculé le point fixe de $p \mapsto e^{-\lambda p}$ mais en fait nous itérons cette fonction.

Explications

Soit p_k la probabilité que la racine ne soit couplée avec aucun de ses enfants dans un arbre tronqué à profondeur k.

- $p_0 = 1$
- $p_1 = e^{\lambda}$
- puis pour $k \geq 0$

$$p_{k+1} = e^{-\lambda p_k}.$$

Nous avons calculé le point fixe de $p \mapsto e^{-\lambda p}$ mais en fait nous itérons cette fonction.

Explications

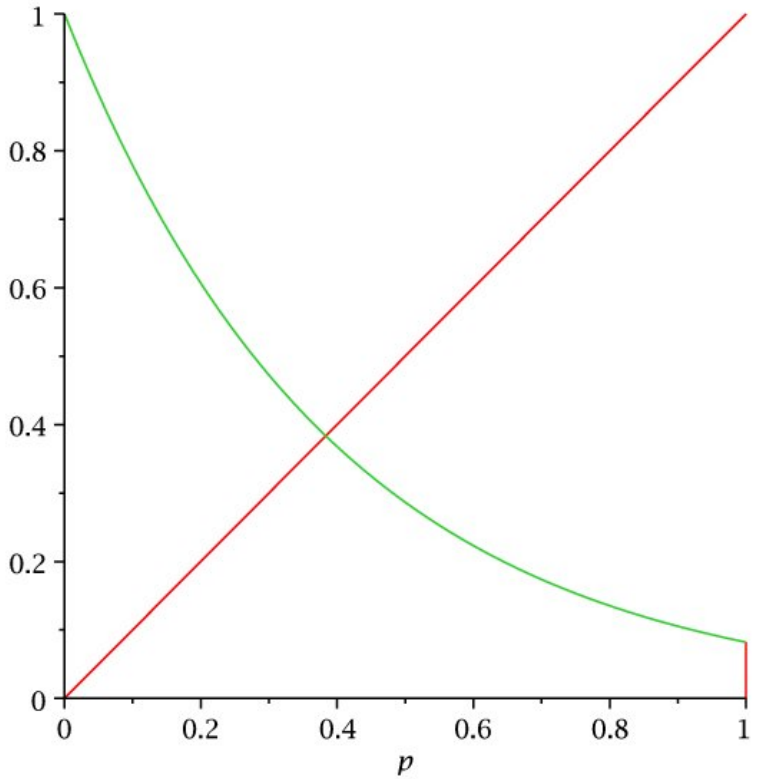
Soit p_k la probabilité que la racine ne soit couplée avec aucun de ses enfants dans un arbre tronqué à profondeur k.

- $p_0 = 1$
- $p_1 = e^{\lambda}$
- puis pour $k \geq 0$

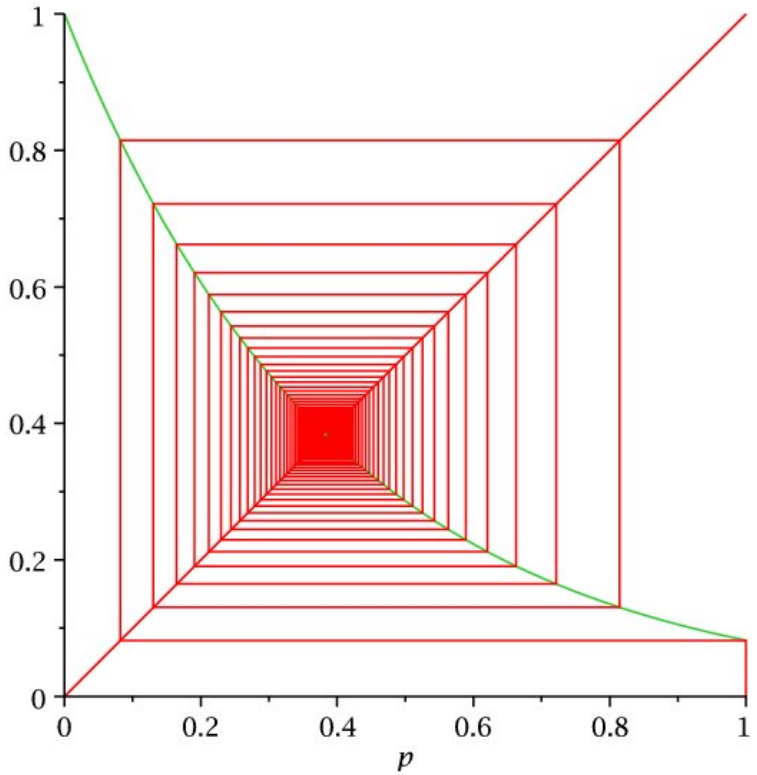
$$p_{k+1} = e^{-\lambda p_k}.$$

Nous avons calculé le point fixe de $p \mapsto e^{-\lambda p}$ mais en fait nous itérons cette fonction.

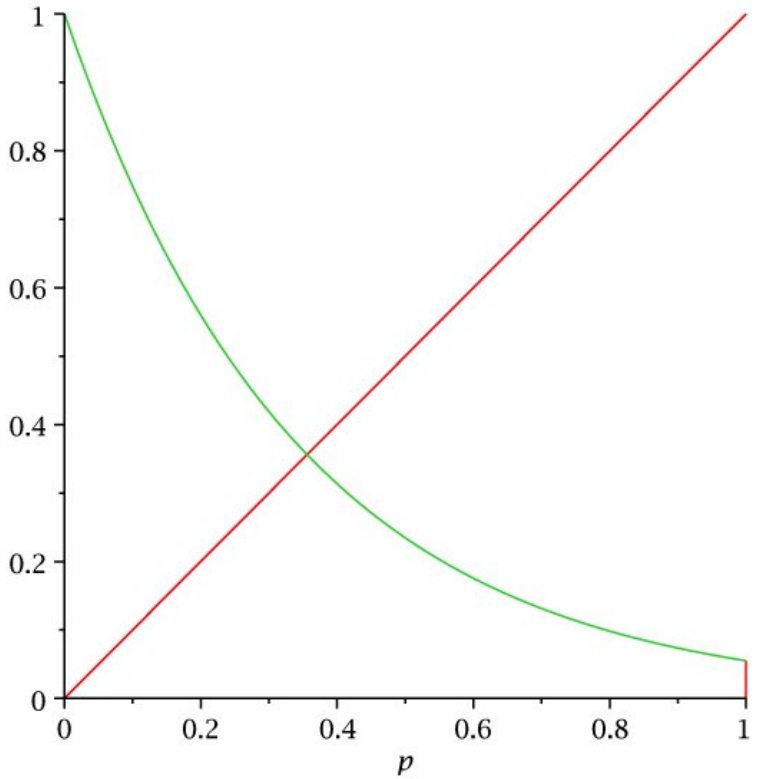
Itérons pour $\lambda < e$



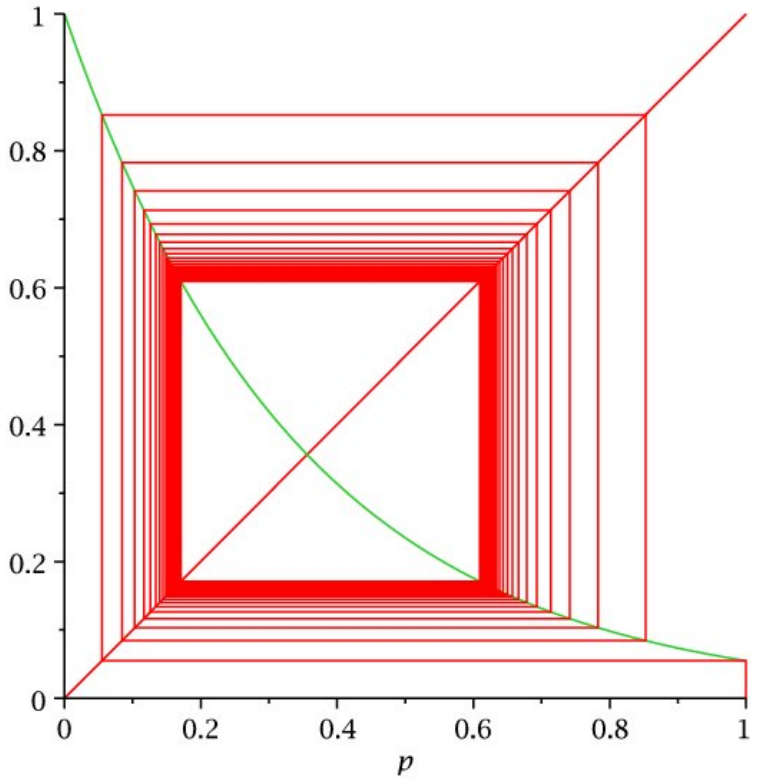
Itérons pour $\lambda < e$



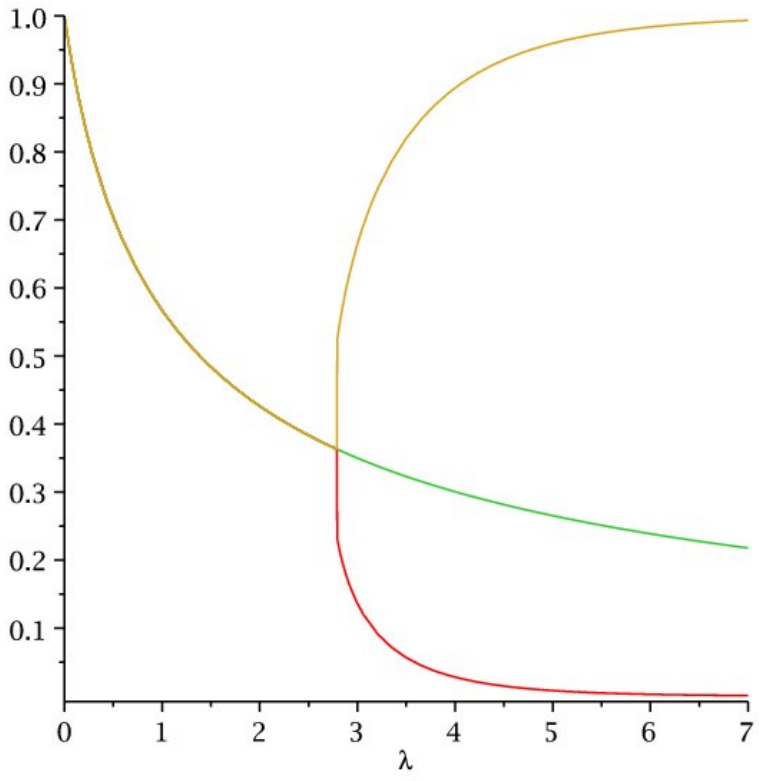
Itérons pour $\lambda > e$



Itérons pour $\lambda > e$



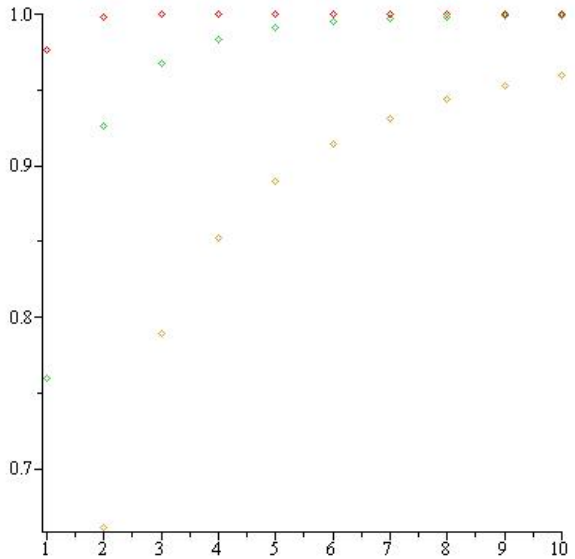
Une transition de phase



Pour $\lambda > e$, les conditions au bord ont une influence sur la racine de l'arbre.

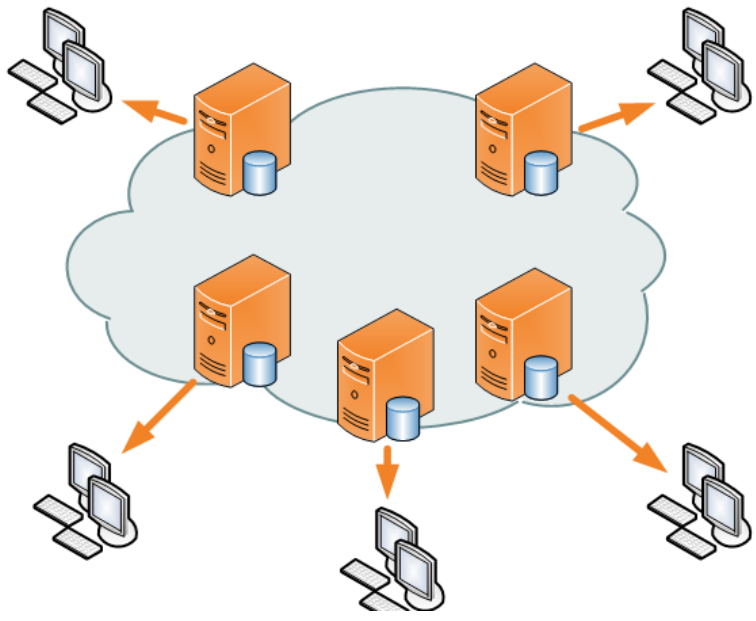
Retour à l'équilibrage de charge

La méthode de la cavité permet de calculer les performances du système.



Charge maximale en fonction de la capacité $k = 1, \ldots, 10$ des boîtes pour un système avec 4 choix pour $\ell = 1, 2, 3$ balles.
Travail de thèse de Justin Salez (Paris 7)

Extension: distributed Content Distribution Network



Algorithme pour une gestion distribuée des caches qui s'adaptent automatiquement à la popularité des contenus.
Travail de thèse de Mathieu Leconte (Technicolor)

Conclusion

- Interactions fructueuses entre mathématique, physique statistique et informatique.
- Champs actif de recherche avec de nombreuses questions ouvertes (aspects algorithmiques) et de nouvelles applications (acquisition compressée de données).

MERCI!

Conclusion

- Interactions fructueuses entre mathématique, physique statistique et informatique.
- Champs actif de recherche avec de nombreuses questions ouvertes (aspects algorithmiques) et de nouvelles applications (acquisition compressée de données).

MERCI!

Conclusion

- Interactions fructueuses entre mathématique, physique statistique et informatique.
- Champs actif de recherche avec de nombreuses questions ouvertes (aspects algorithmiques) et de nouvelles applications (acquisition compressée de données).

MERCI!