Bases de données documentaires et distribuées
Cours NFE04
L'algorithme PageRank

Auteurs : Raphaël Fournier-S'niehotta, Philippe Rigaux, Nicolas Travers
prénom.nom@cnam.fr

Département d'informatique
Conservatoire National des Arts & Métiers, Paris, France

Plan du cours

PageRank : importance déduite du graphe des documents

Dans le cas du Web (et quelques autres systèmes), les documents sont liés par des hyperliens.

La structure de la collection est donc celle d'un **graphe orienté**.

Intuition

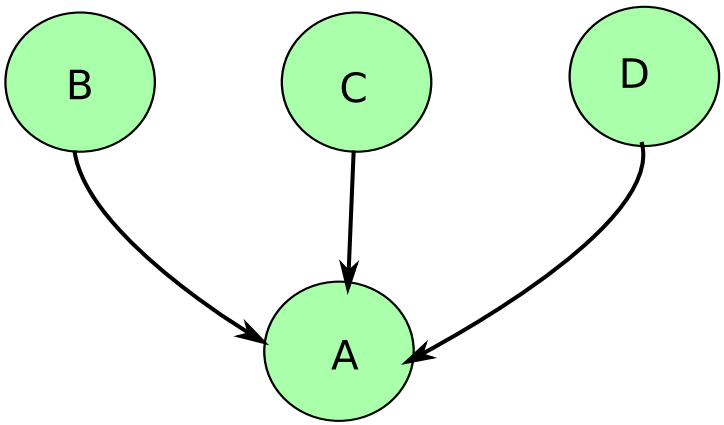
Un document vers lequel convergent beaucoup de chemins est un document **important**.

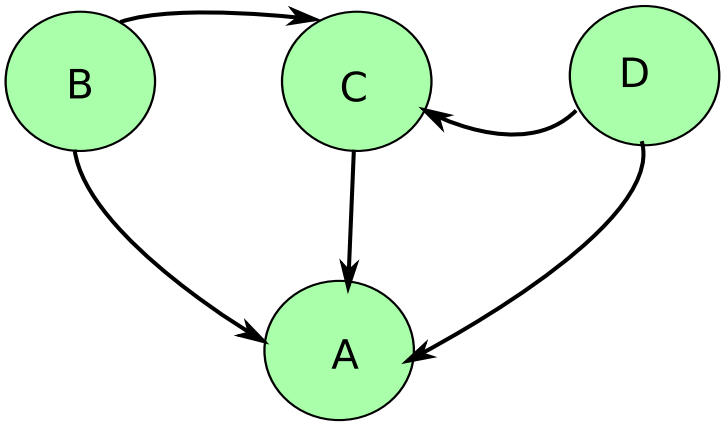
En combinant avec des mesures de pertinence (tf/idf), on obtient un moyen d'améliorer le classement.

PageRank : définition et exemples

Définition

L'indicateur *PageRank* (PR) d'une page p_i est la **probabilité** qu'un utilisateur suivant les liens de manière aléatoire arrive sur P_i.



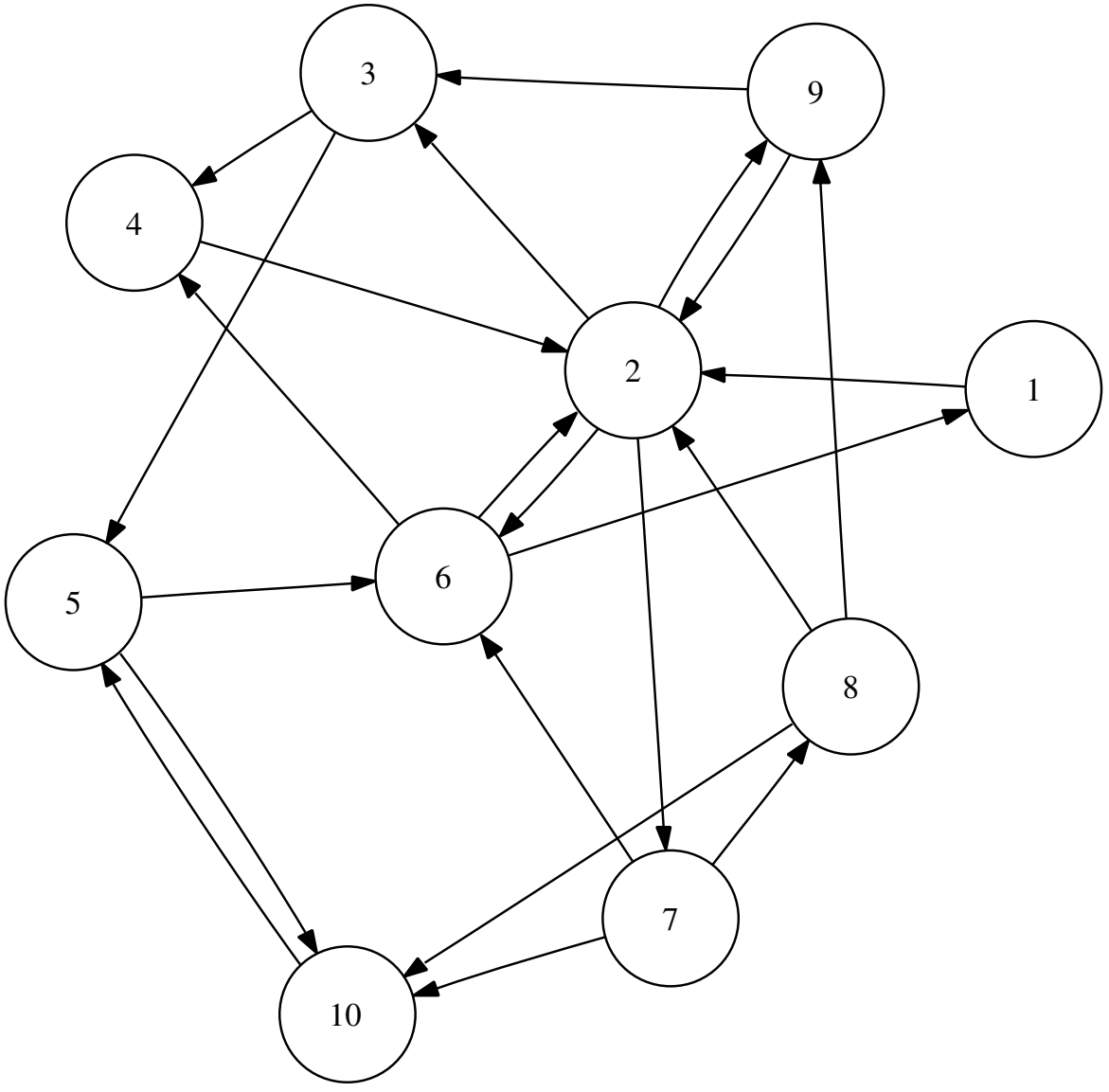


À gauche : la probabilité d'arriver en A en **une** étape est
$PR(A) = PR(B) + PR(C) + PR(D)$

À droite ?

Au départ, chaque page a un PR de 0,25. Quel est le PR après une itération ?
Et après deux ?

Un exemple plus complet



On construit une **matrice de transition**

$$\begin{cases} g_{ij} = 0 & \text{s'il n'y a pas de lien entre les pages } i \text{ et } j\,; \\ g_{ij} = \frac{1}{n_i} & \text{sinon, } n_i \text{ étant le nombre de liens sortant de la page } i. \end{cases}$$

$$G = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Calcul du PageRank, pas à pas

Je veux calculer la probablité d'être en N_2 à l'étape e. J'ai besoin :

- de la probabilité d'être sur chaque nœud N_i à l'étape $e-1$
  $\Rightarrow$ c'est le vecteur des PageRank, appelons-le v.
- de la probabilité d'arriver au nœud N_2 venant du nœud N_i
  $\Rightarrow$ c'est la seconde **colonne** de la matrice.

Allons-y. Au départ, le vecteur des PageRank est uniforme

$$v = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]$$

La seconde colonne de la matrice, **transposée** est :

$$C_2 = [1, 0, 0, 1, 0, 1/3, 0, 1/3, 1/2, 0]$$

Ce qui donne la probabilité d'arriver en N_2 à la première itération

$$0.1 \times 1 + 0.1 \times 1 + 0.1 \times 1/3 + 0.1 \times 1/3 + 0.1 \times 1/2 = 0.317$$

Interprétation : j'ai 10% de chances d'être en N_1, 100% de chances, étant en N_1, d'aller en N_2, etc.

Calcul du PageRank, généralisé

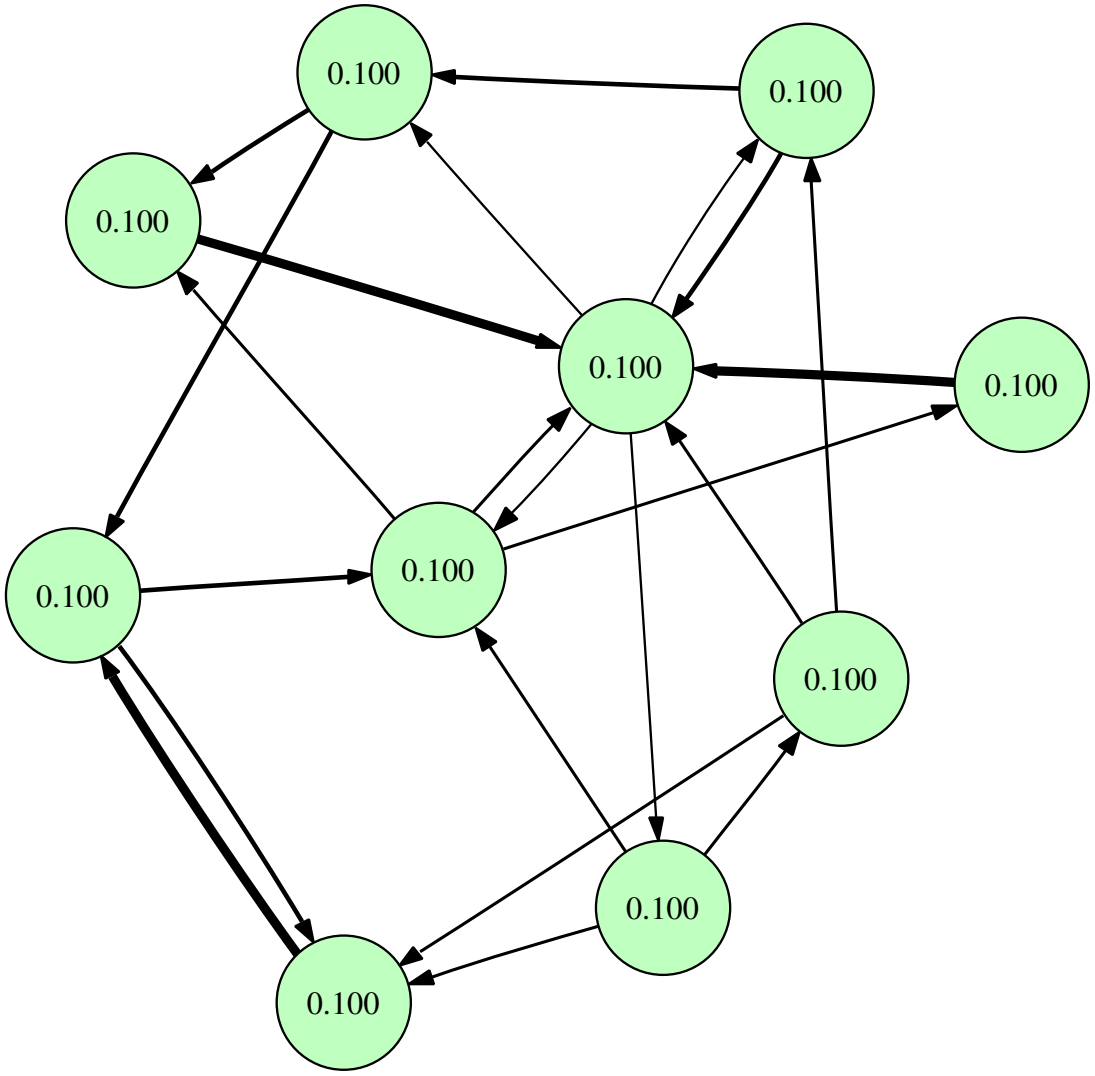
On effectue le calcul précédent pour tous les nœuds, et autant de fois que nécessaire.

- On construit par itérations un vecteur contenant l'indicateur PR de chaque page du graphe.
  Appelons-le v ; il contient autant de coordonnées que de pages du Web...
- v est initialisé avec une distribution uniforme ($v[i] = \frac{1}{|v|}$).
  Sur notre exemple, la valeur initiale est 1/10.
- À chaque itération, on ajuste v on calculant la probabilité qu'un déplacement amène sur chaque nœud.
  On multiplie le vecteur v par la **transposée** de G (les colonnes donnent la probabilité d'**arriver** sur un nœud).

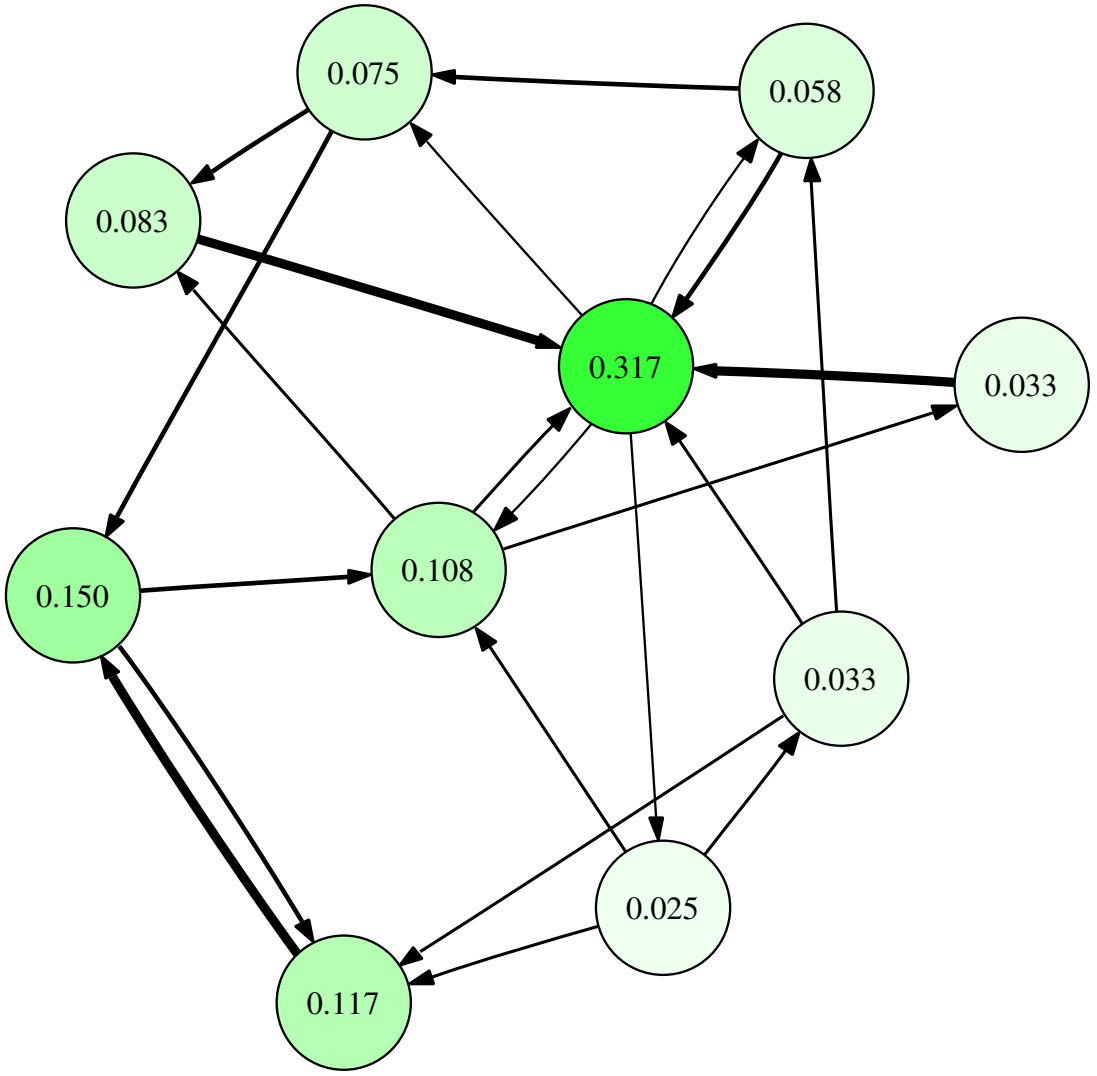
On peut montrer qu'il y a **convergence** du vecteur v vers une limite.

$$\mathrm{pr}(i) = \left(\lim_{k \to +\infty} (G^T)^k v \right)_i$$

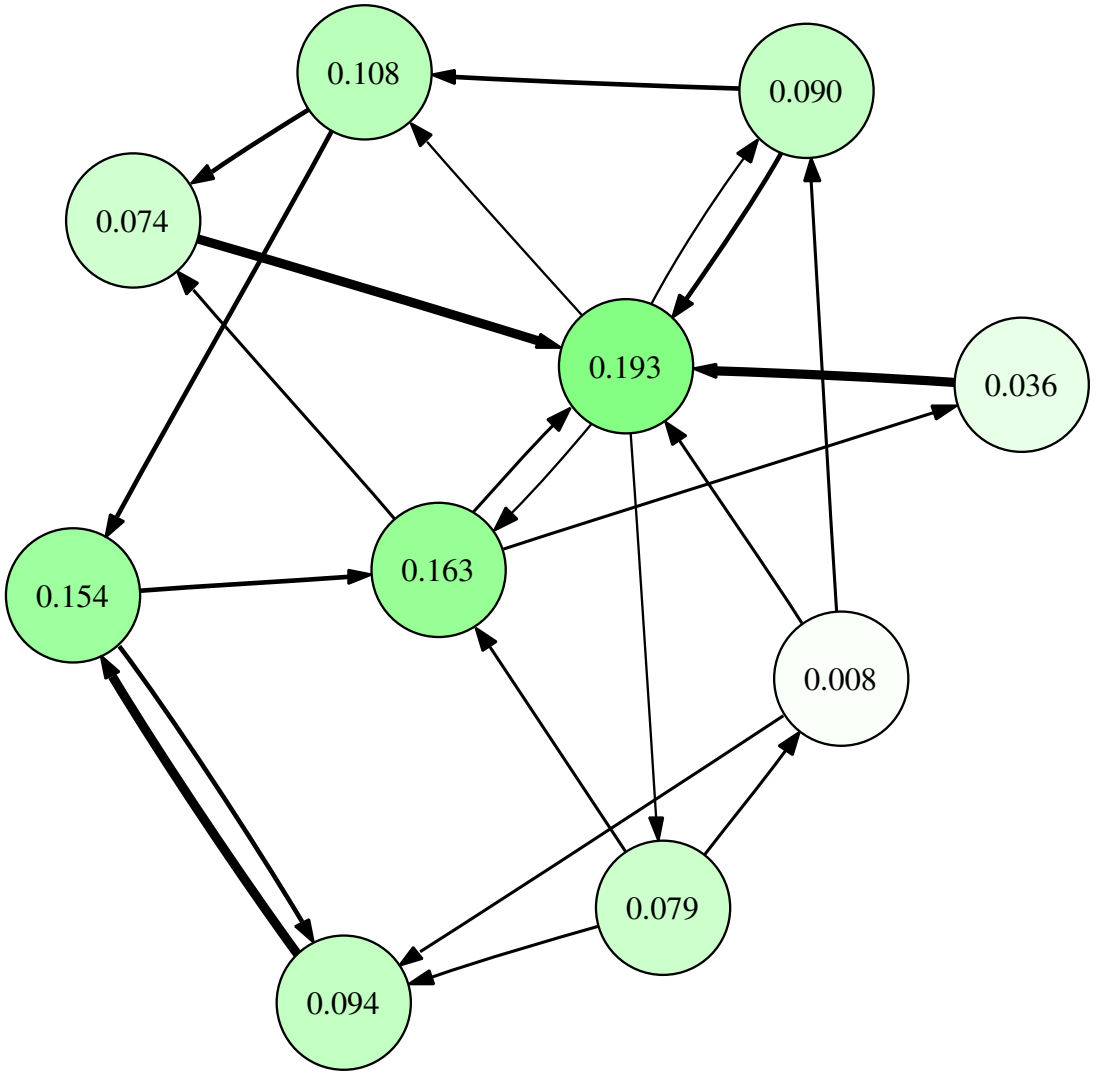
Queques itérations PageRank



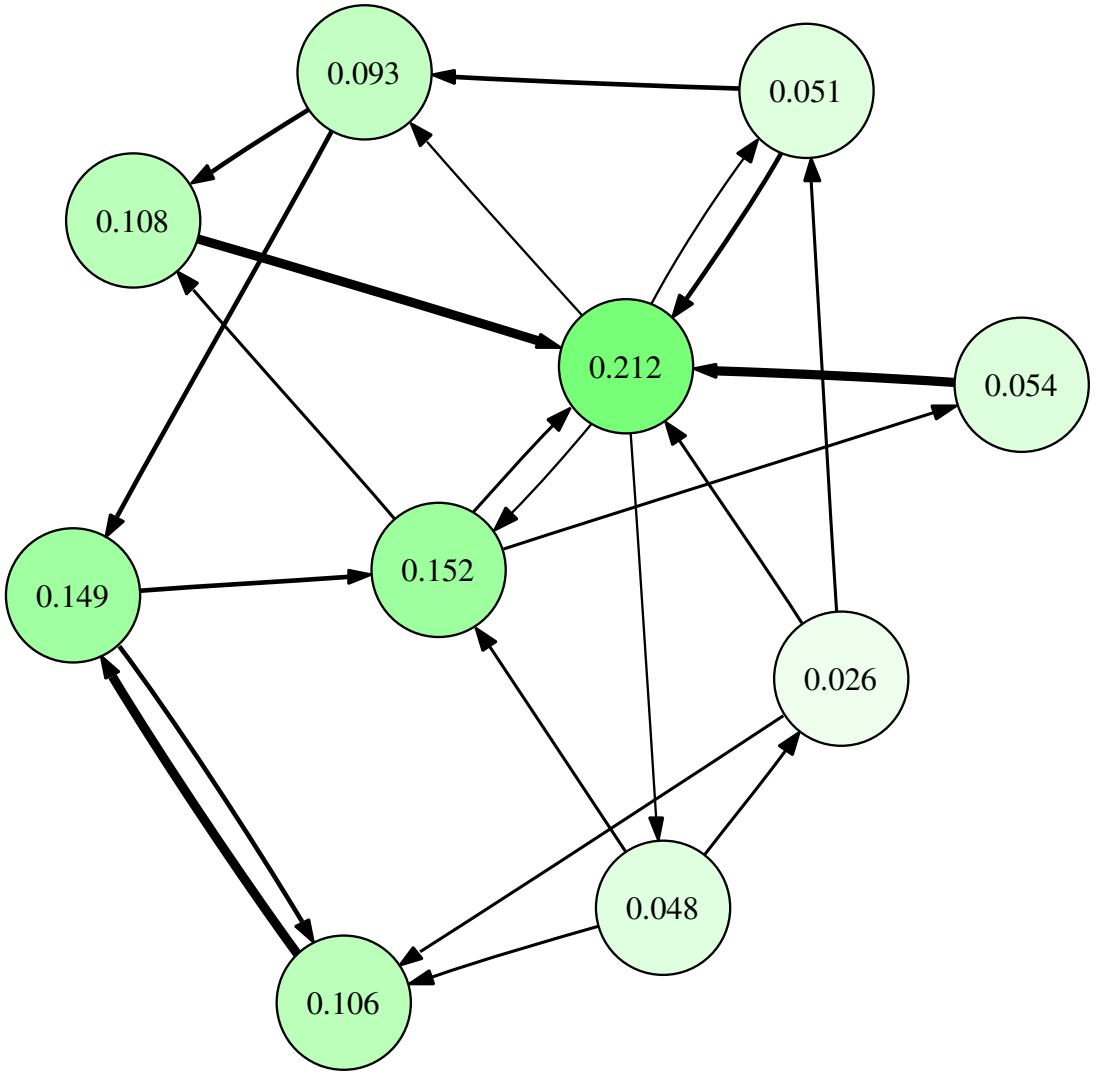
Queques itérations PageRank



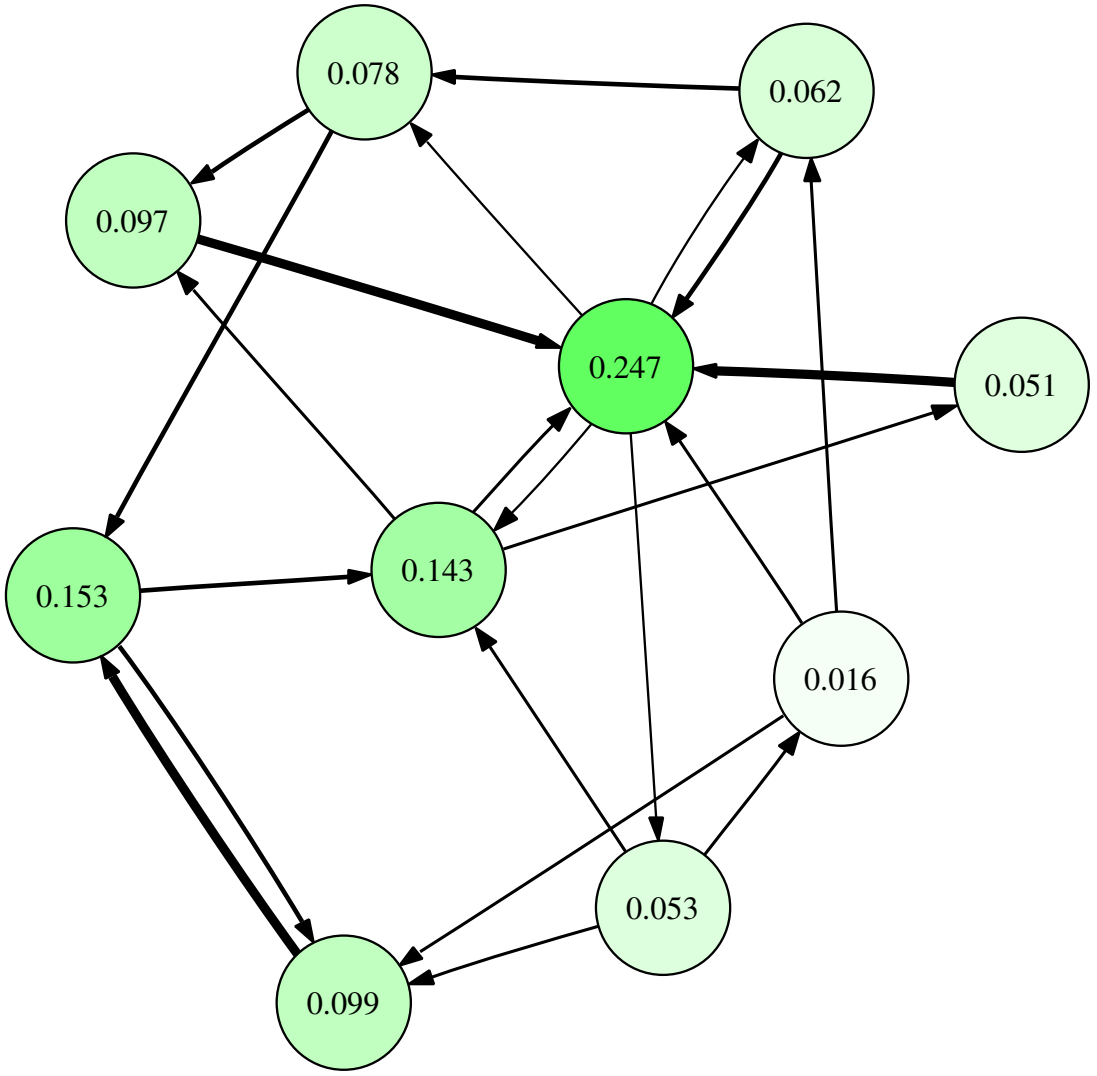
Queques itérations PageRank



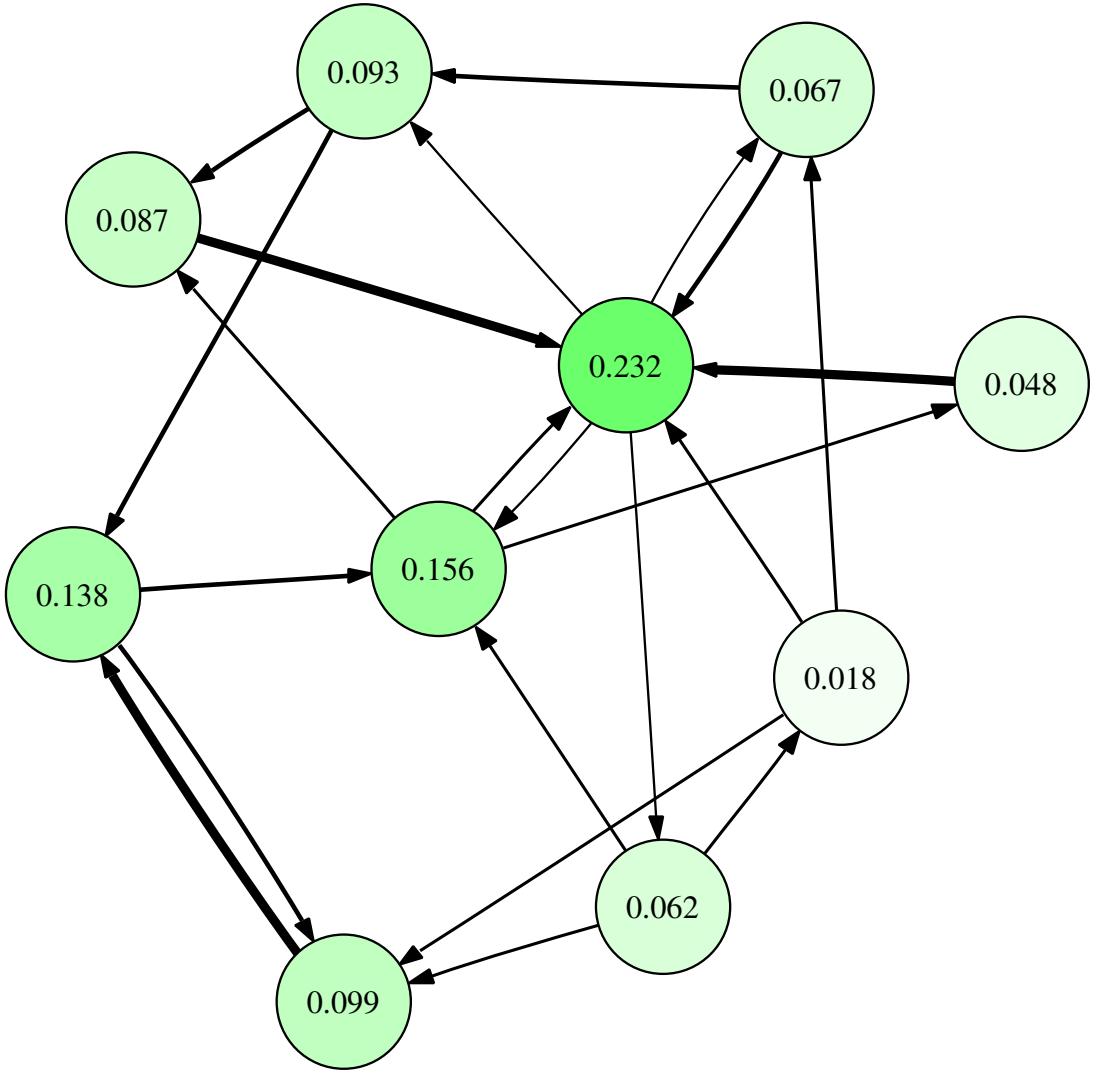
Queques itérations PageRank



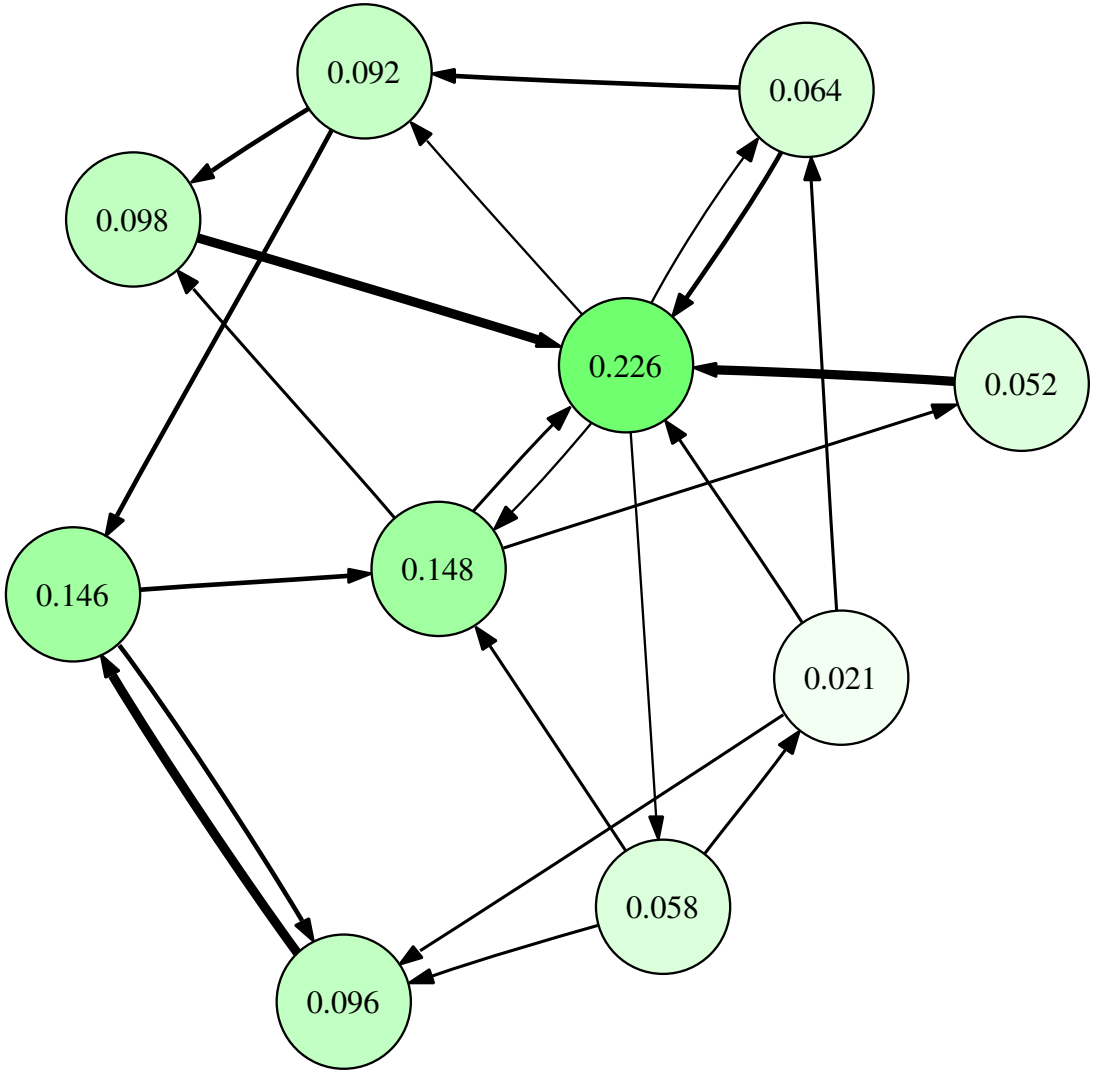
Queques itérations PageRank



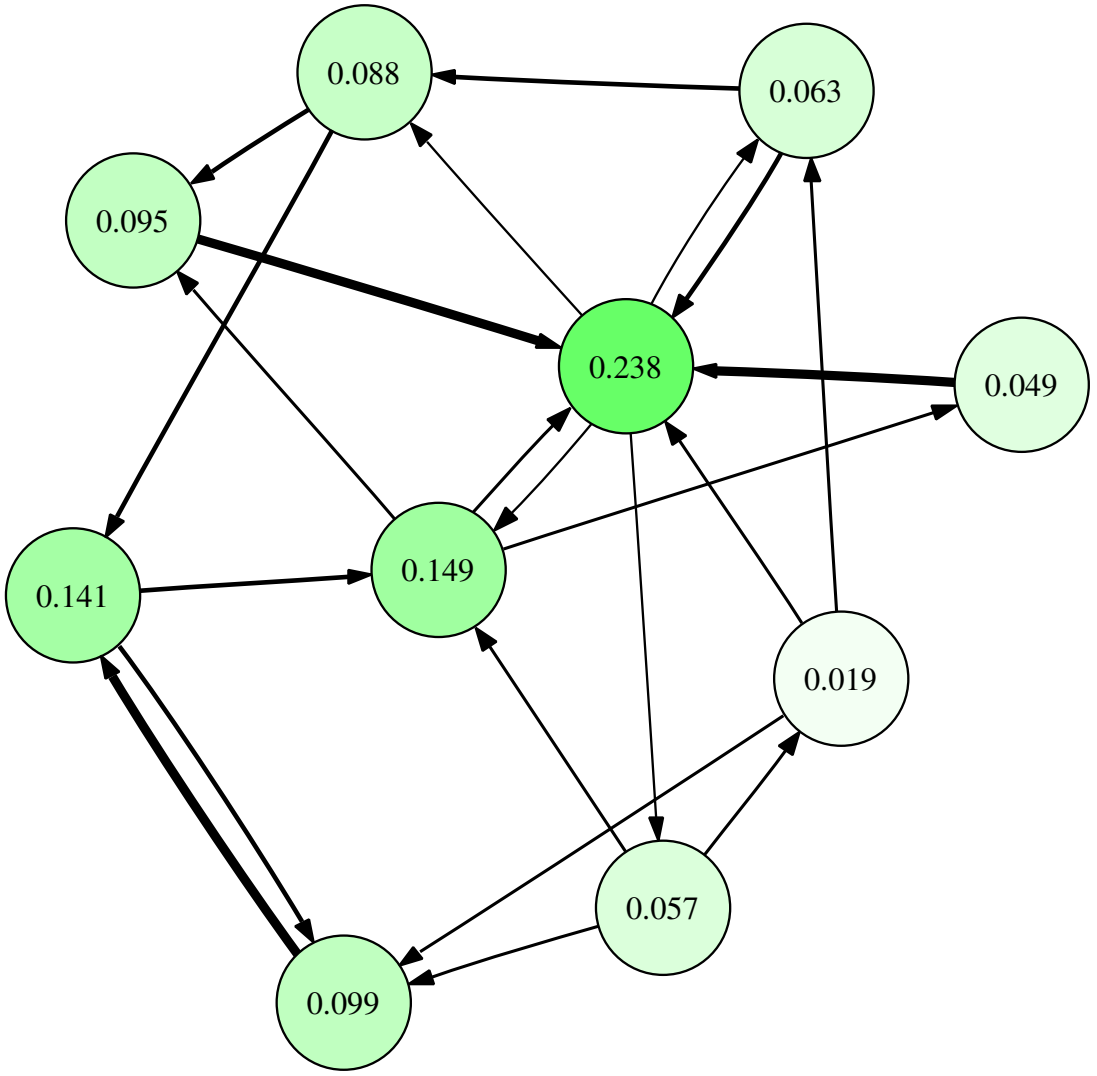
Queques itérations PageRank



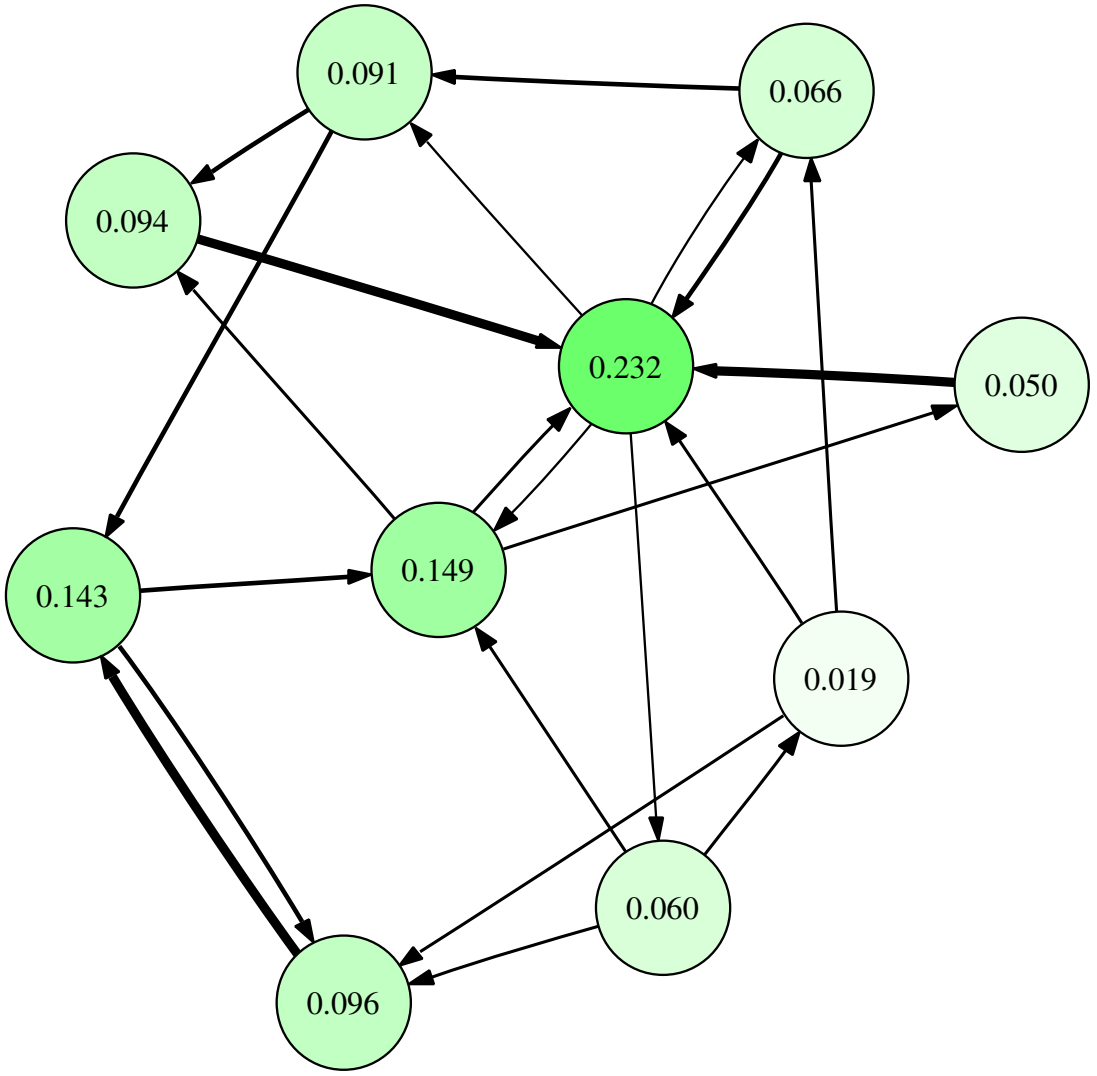
Queques itérations PageRank



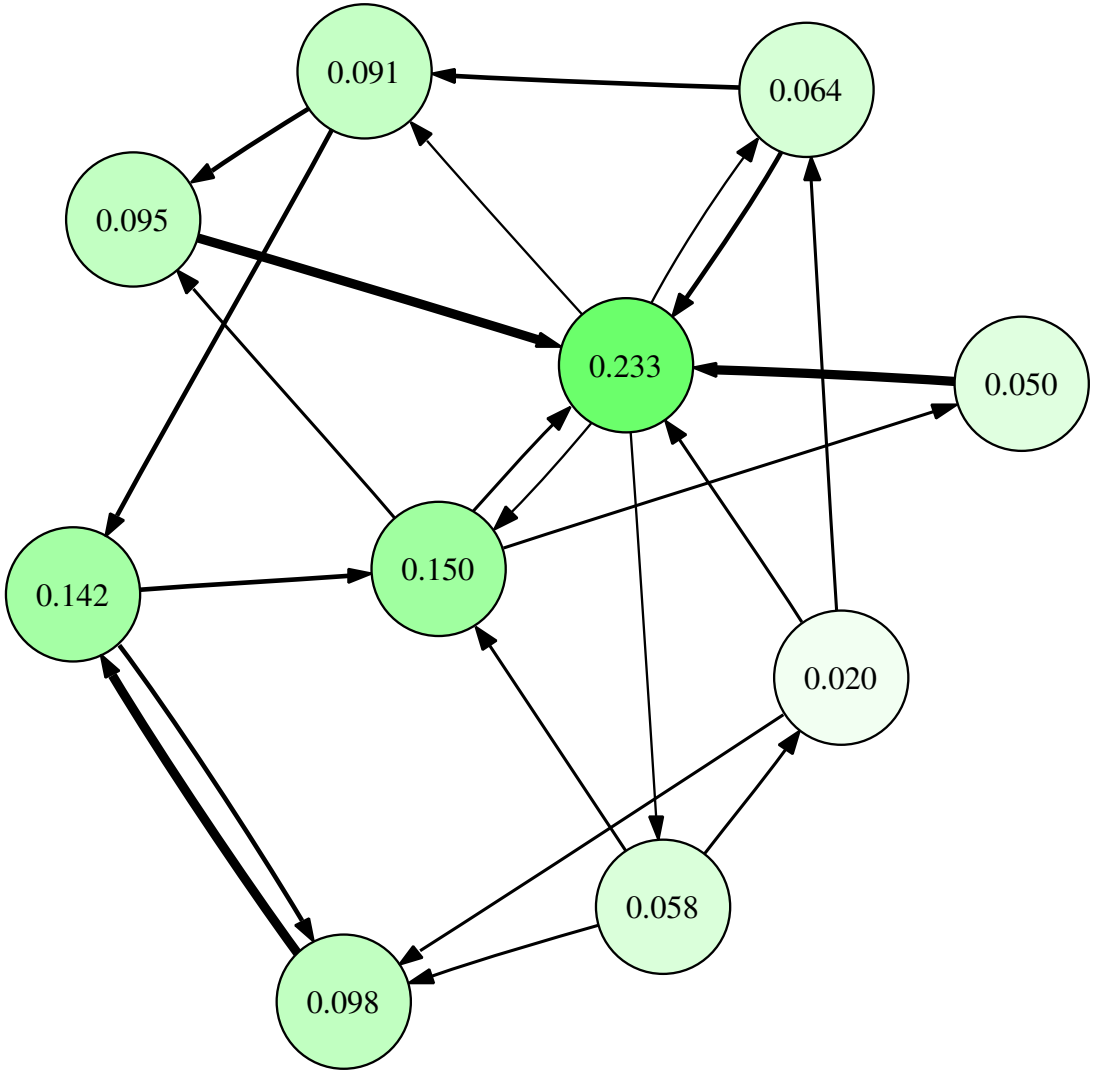
Queques itérations PageRank



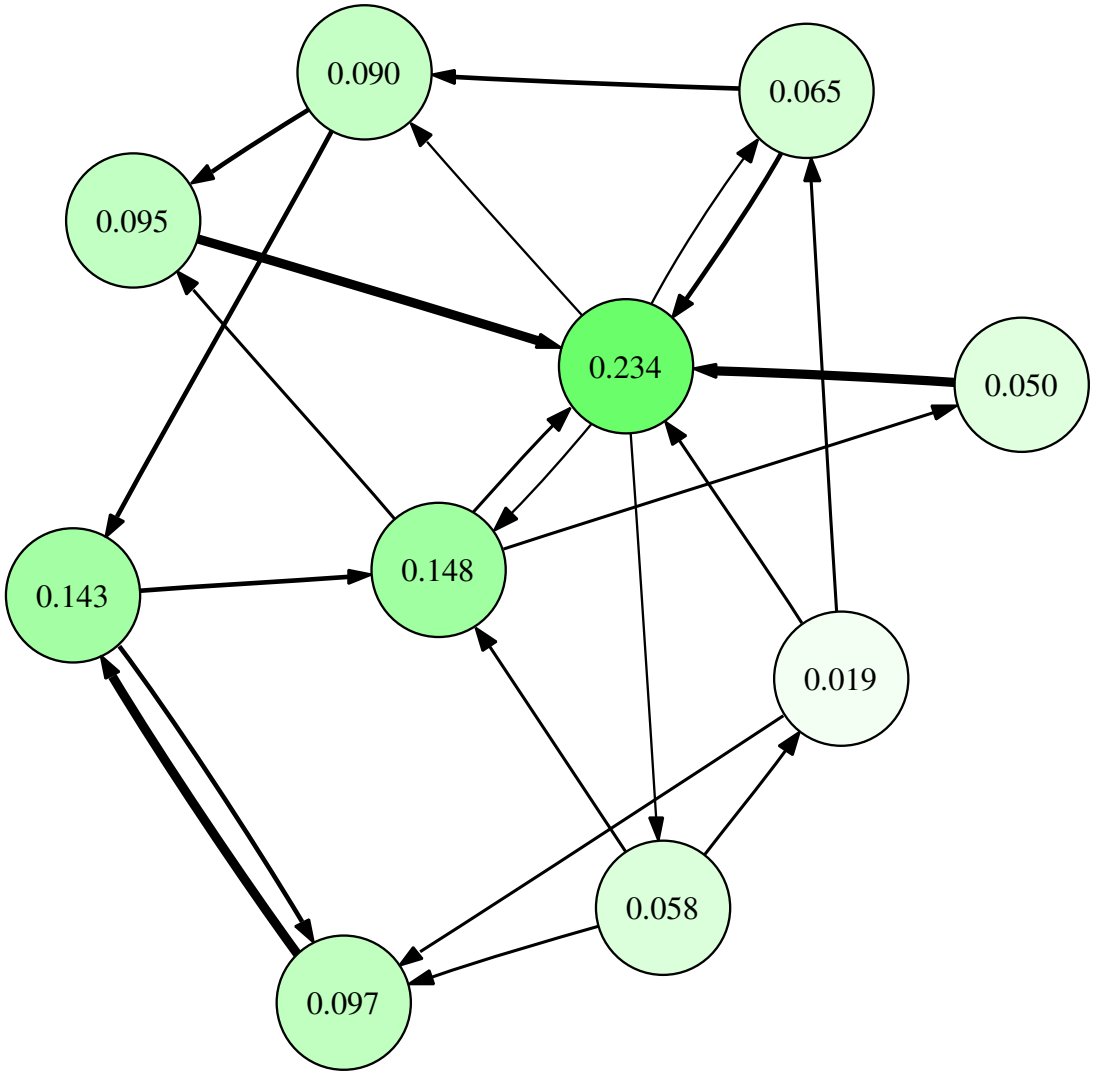
Queques itérations PageRank



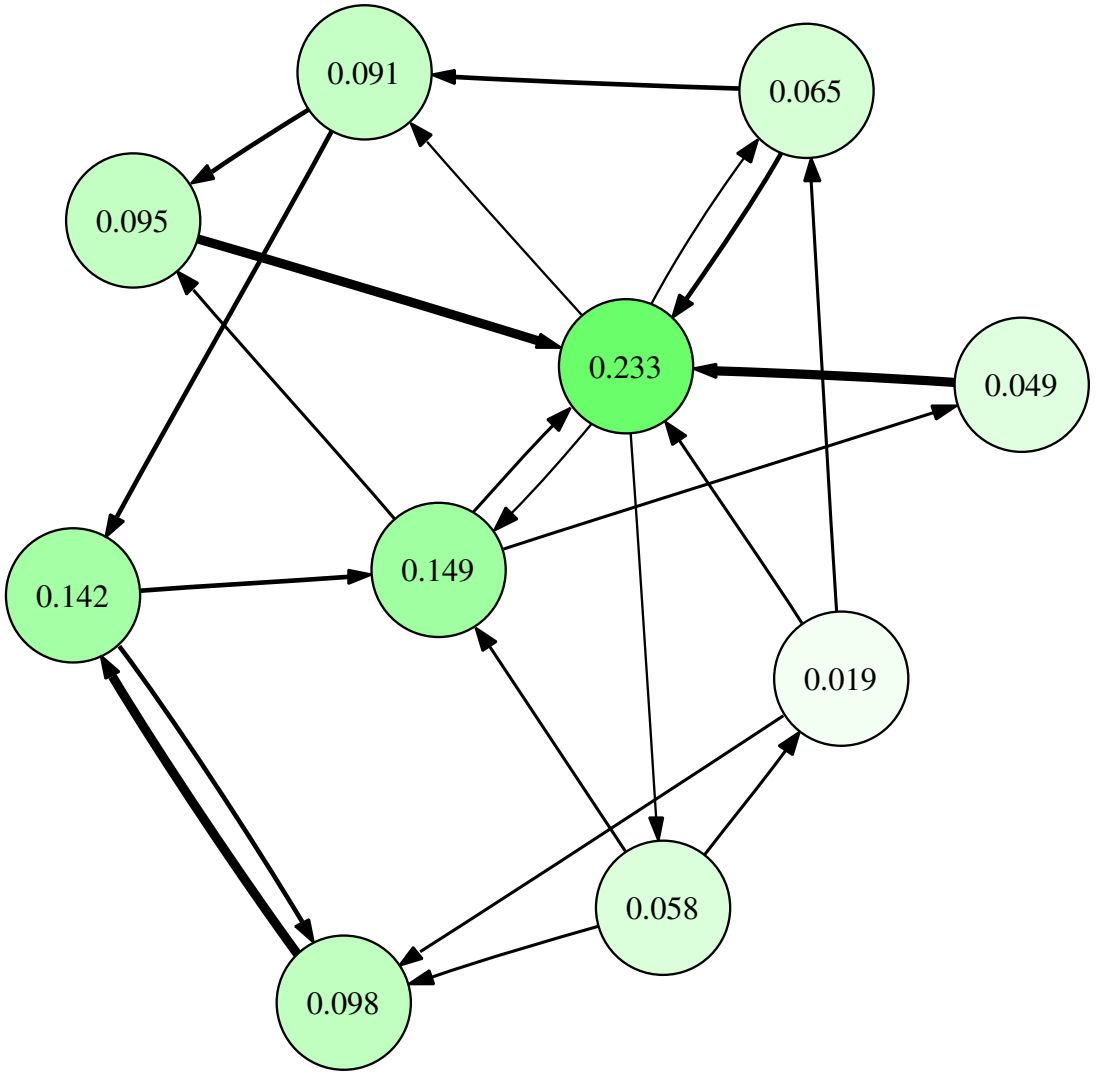
Queques itérations PageRank



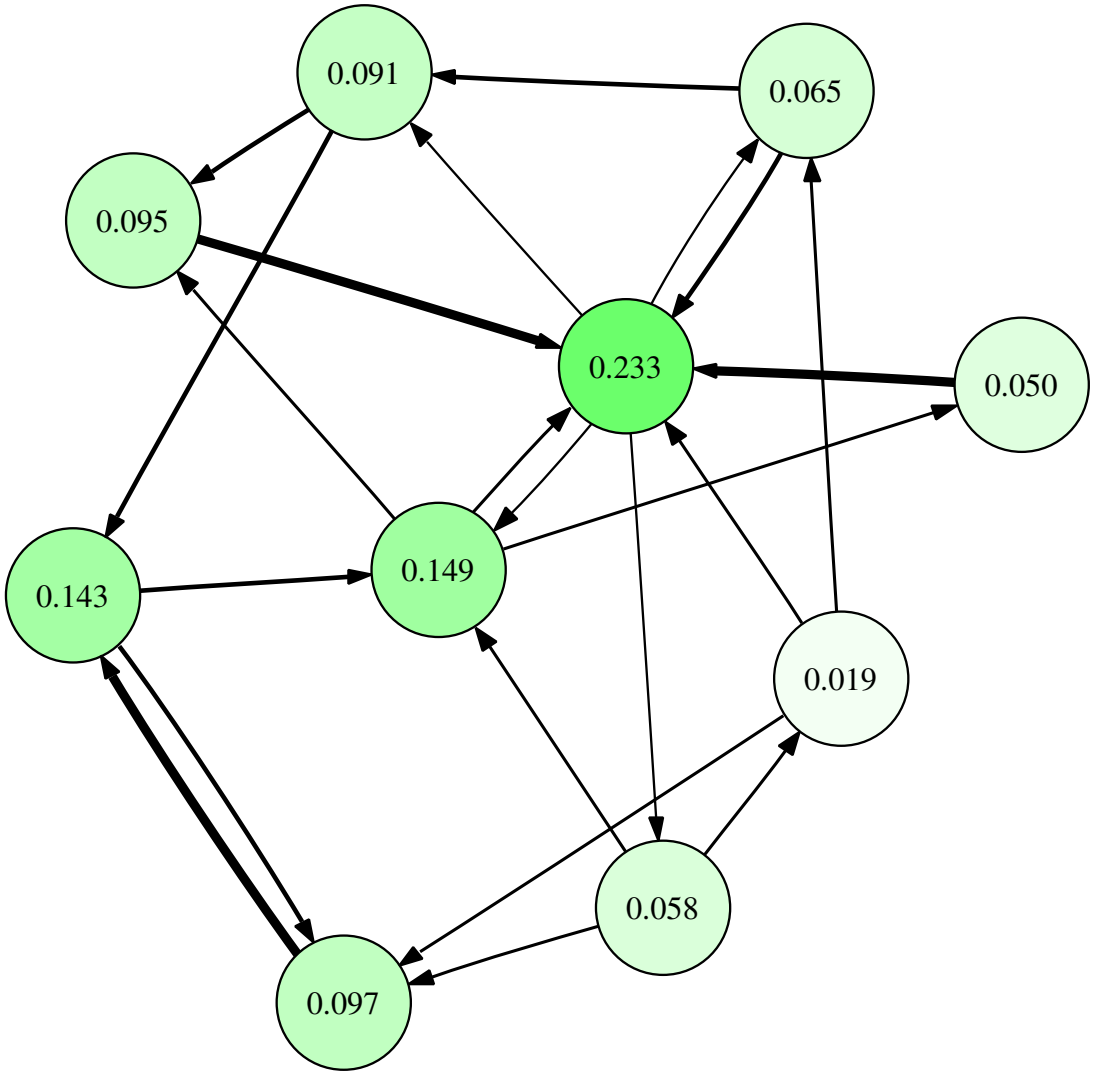
Queques itérations PageRank



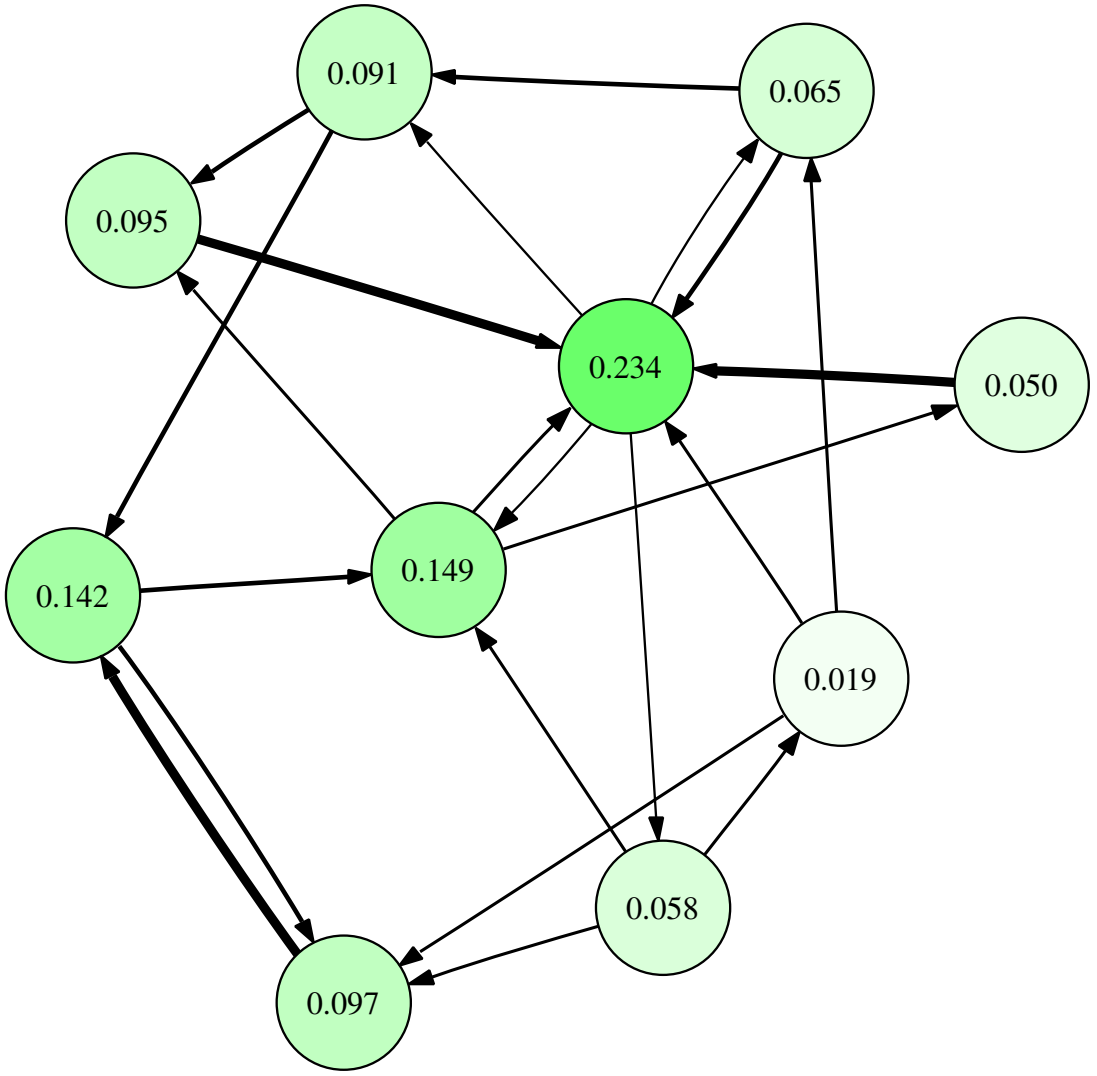
Queques itérations PageRank



Queques itérations PageRank



Queques itérations PageRank



Petite extension pratique

Pour mieux modéliser le comportement d'un utilisateur, on s'autorise des **sauts** d'une page à une autre, sans qu'il y ait nécessairement de lien.

À chaque étape, on prend en compte la possibilité d'un tel saut avec une probabilité d ($1-d$: **damping factor**). Ce qui donne :

$$\mathrm{pr}(i) = \left(\lim_{k \to +\infty} ((1-d)G^T + dU)^k v \right)_i$$

où U est une matrice contenant $\frac{1}{N}$ dans chaque cellule.