Modélisation et propriétés des graphes

Raphaël Fournier-S'niehotta

CNAM Paris, fournier@cnam.fr

HTT-FOD
RCP216
2021-2022

le cnam

Plan

Propriétés des graphes

Introduction : pourquoi des indicateurs ?

Dans un graphe, on cherche souvent :

- qui sont les figures centrales (individus) ?
  - centralité(s)
- quelles sont les interactions fréquentes entre personnes ?
  - transitivité, réciprocité
  - équilibre, statut
- quels sont les utilisateurs similaires
  - similarité

Il nous faut des manières d'estimer certains phénomènes, pour répondre à ces questions (et d'autres)

Centralités

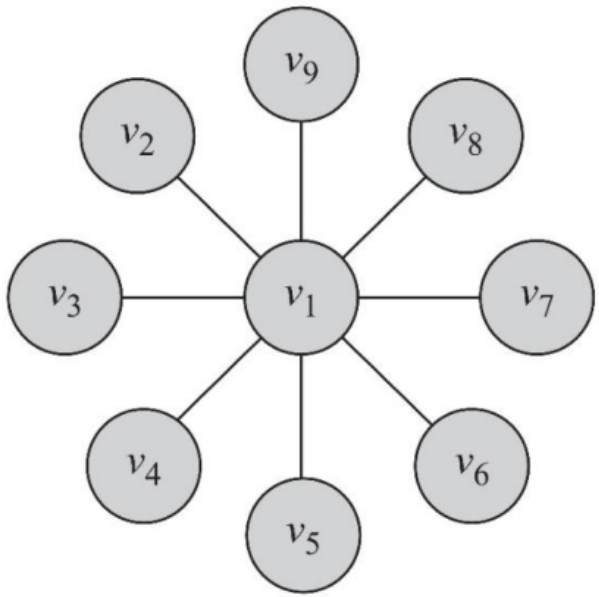
- Importance des voisins
- Interconnexions
- Atteindre les autres

Centralité de degré

- C'est un indicateur égal au degré du nœud :

$$C_d(v_i) = d_i$$

- exemple : $C(v_1) = 8$, $C(v_j) = 1, \forall j \neq 1$



Dans un graphe orienté, on peut utiliser le degré entrant, sortant, ou une combinaison des deux:

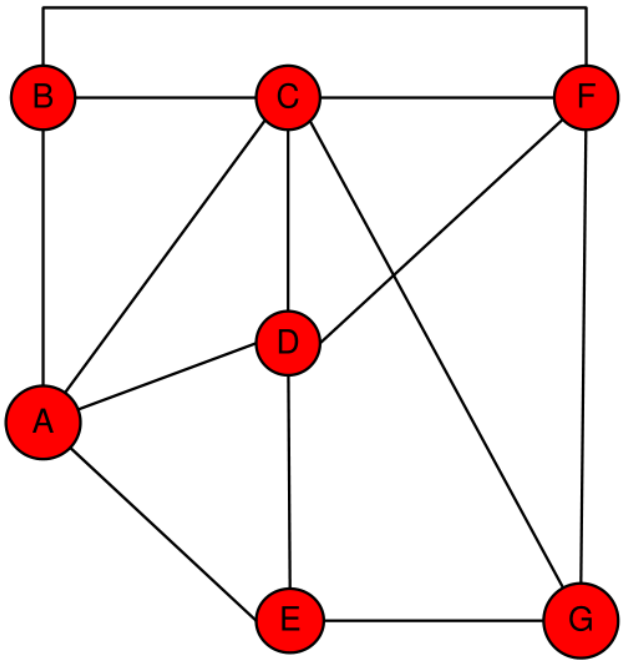
- $C(v_i) = d_i^{in}$ (prestige)
- $C(v_i) = d_i^{out}$ (grégarisme)
- $C(v_i) = d_i^{in} + d_i^{out}$

Centralité de degré normalisée

On peut normaliser la centralité de degré :

- par le degré maximum *possible* : $C_d^{norm} = \frac{d_i}{n-1}$
- par le degré maximum $C_d^{max} = \frac{d_i}{\max_j d_j}$
- par la somme des degrés $C_d^{sum} = \frac{d_i}{\sum_j d_j} = \frac{d_i}{2m}$

Exemple de centralité



Node	Degree	Centrality	Rank
A	4	2/3	**2**
B	3	1/2	**5**
C	5	5/6	**1**
D	4	2/3	**2**
E	3	1/2	**5**
F	4	2/3	**2**
G	3	1/2	**5**

Centralité de vecteur propre (eigenvector)

- Avoir le plus de liens n'est pas, en soi, une garantie de l'importance du nœud (de la personne)
- Avoir des **amis importants** peut être plus significatif
- la centralité de valeur propre généralise la centralité de degré, en incorporant l'importance des voisins (graphe non orienté)
- pour les graphes orientés, on utilise les arêtes entrantes ou sortantes

Formalisation

- On cherche la centralité (inconnue) d'un nœud v, notée x_v
- On souhaite que cette quantité soit élevée quand des nœuds importants pointent vers v (donc des nœuds tels que leur x_i soit grand)
- On peut essayer avec :

$$x_v = \sum_{u=1}^{n} A_{u,v} x_u$$

- Que l'on peut normaliser, pour borner la somme :

$$x_v = \frac{1}{\lambda} \sum_{u=1}^{n} A_{u,v} x_u, \text{avec} \lambda \text{constante}$$

- En notation matricielle :

$$\boldsymbol{A}.\boldsymbol{x} = \lambda \boldsymbol{x}$$

- Donc $\boldsymbol{x}$ est un vecteur propre de la matrice d'adjacence A, correspondant à la valeur propre λ.

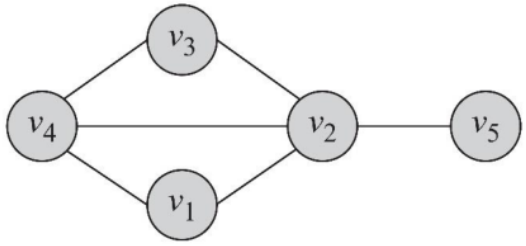
Formalisation (suite)

- Quelle valeur propre ?
- Si l'on impose que les coefficients du vecteur propre doivent être positifs, alors le théorème de Perron-Frobenius assure que seule **la plus grande valeur propre** fournit la mesure de centralité souhaitée
- La composante d'indice v du vecteur propre correspondant donne alors le score de centralité du sommet *j*

Remarques

- les centralités PageRank et Katz sont construites autour de cette idée

Exemple



$$A = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{bmatrix} \Rightarrow \lambda = (2.68,\ -1.74, -1.27,\ 0.33, 0.00)$$

$$\lambda_{max} = 2.68 \Rightarrow C_e = \begin{bmatrix} 0.4119 \\ 0.5825 \\ 0.4119 \\ 0.5237 \\ 0.2169 \end{bmatrix}$$

PageRank

- 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.
- Larry Page, Sergey Brin, 1998
- En combinant avec des mesures de pertinence (tf/idf), on obtient un moyen d'améliorer le classement.

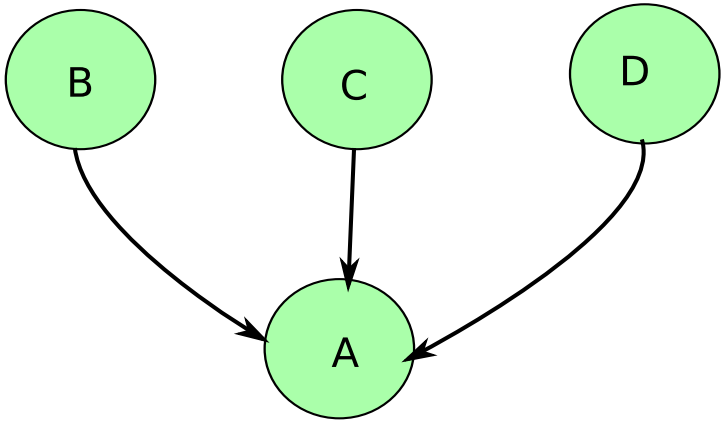
Intuition

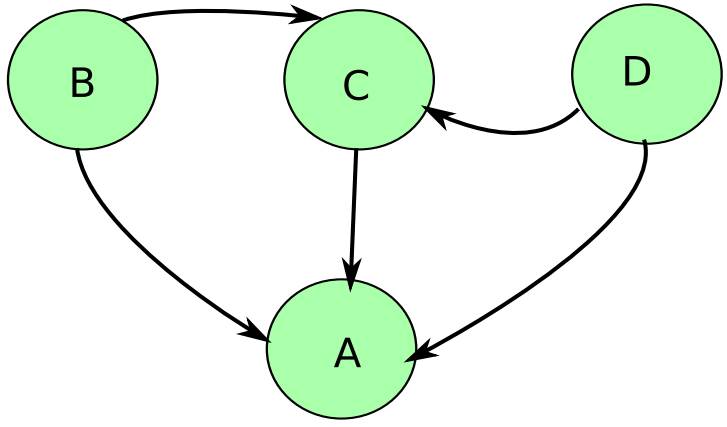
Un document vers lequel convergent beaucoup de chemins est un document important.

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.



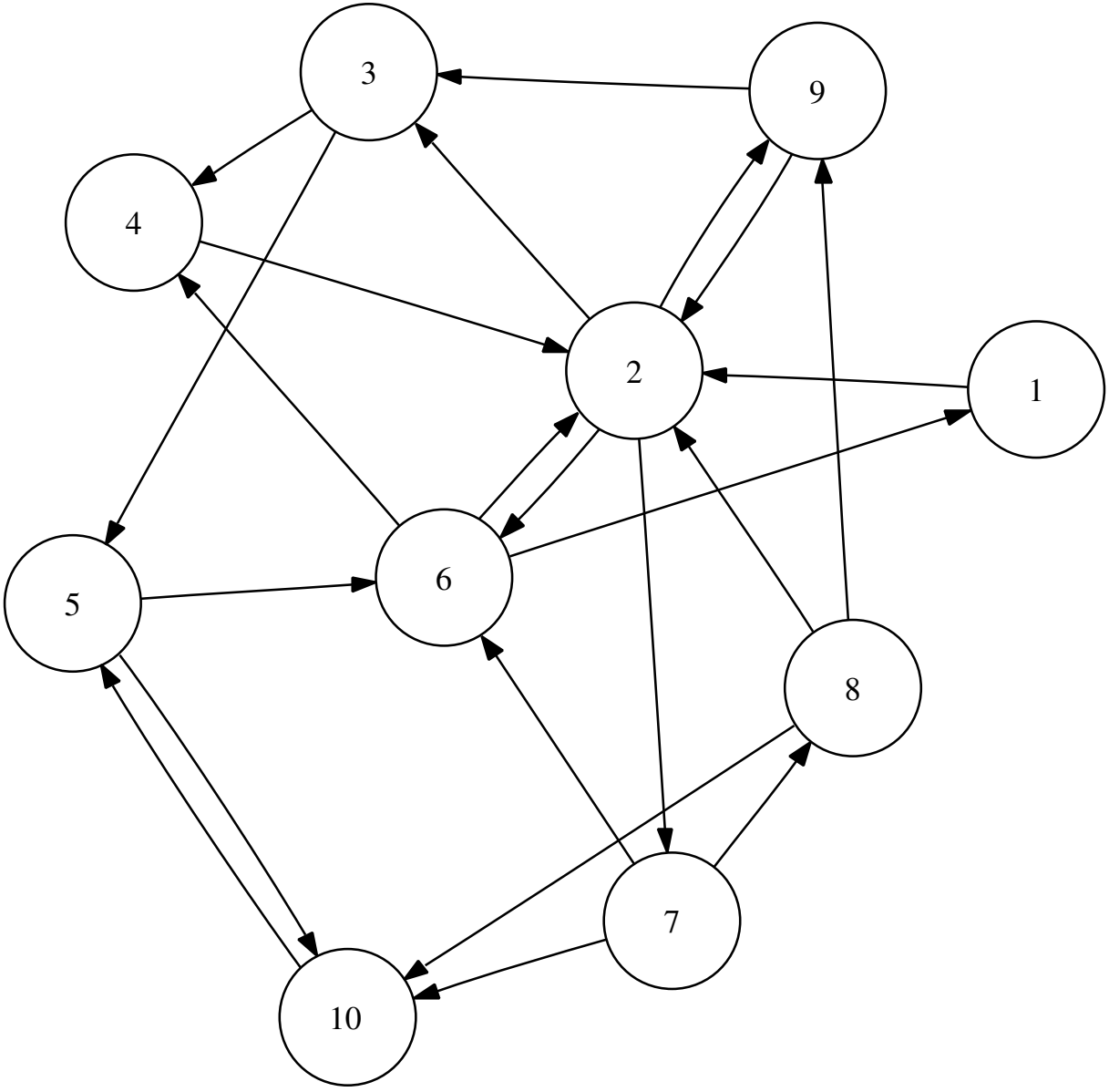


À gauche : la probabilité d'arriver en A en une étape est $PR(A) = PR(B) + PR(C) + PR(D)$

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

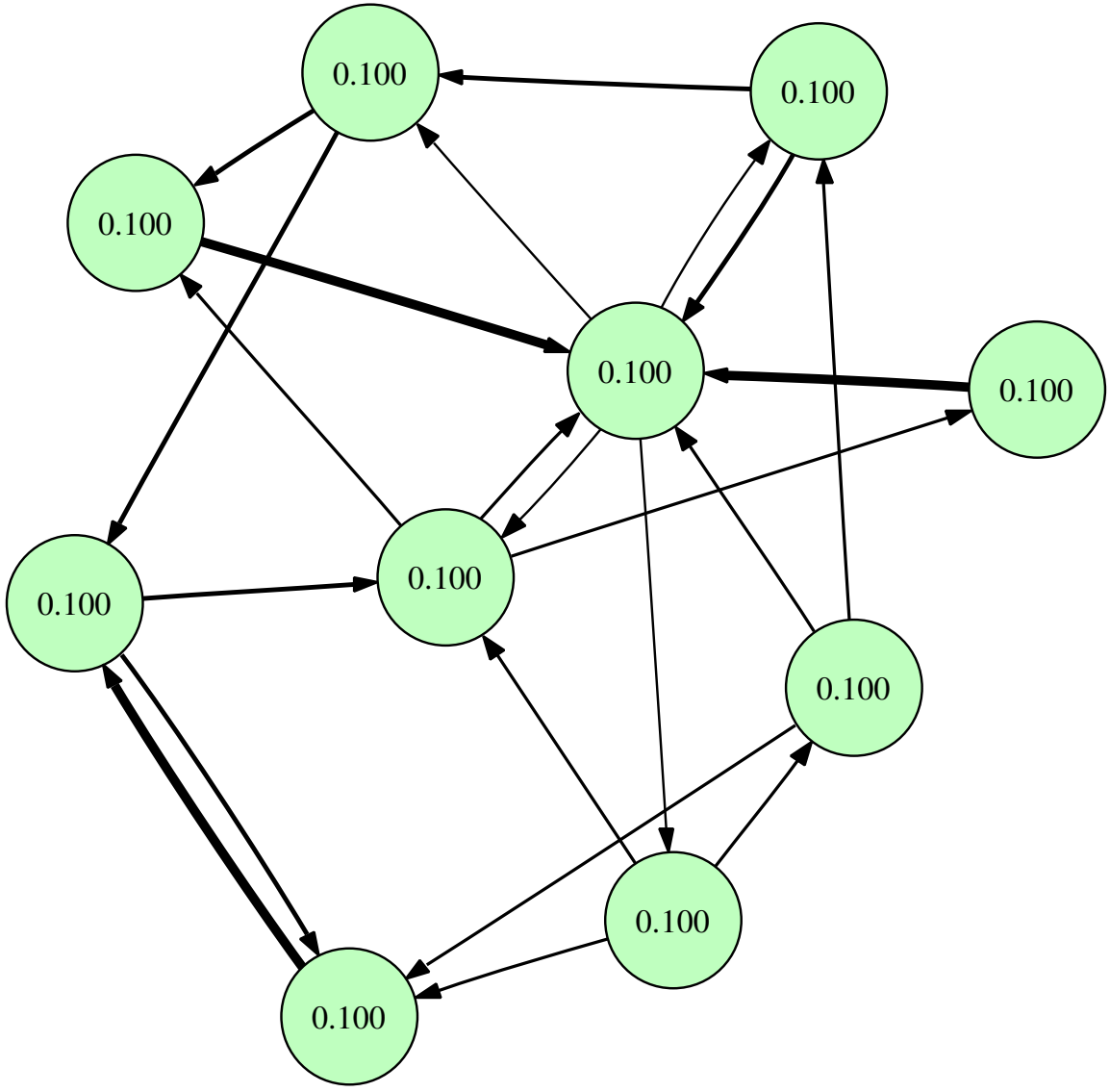
On effectue le calcul précédent pour tous les nœuds, et autant de fois que nécessaire.

- 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).

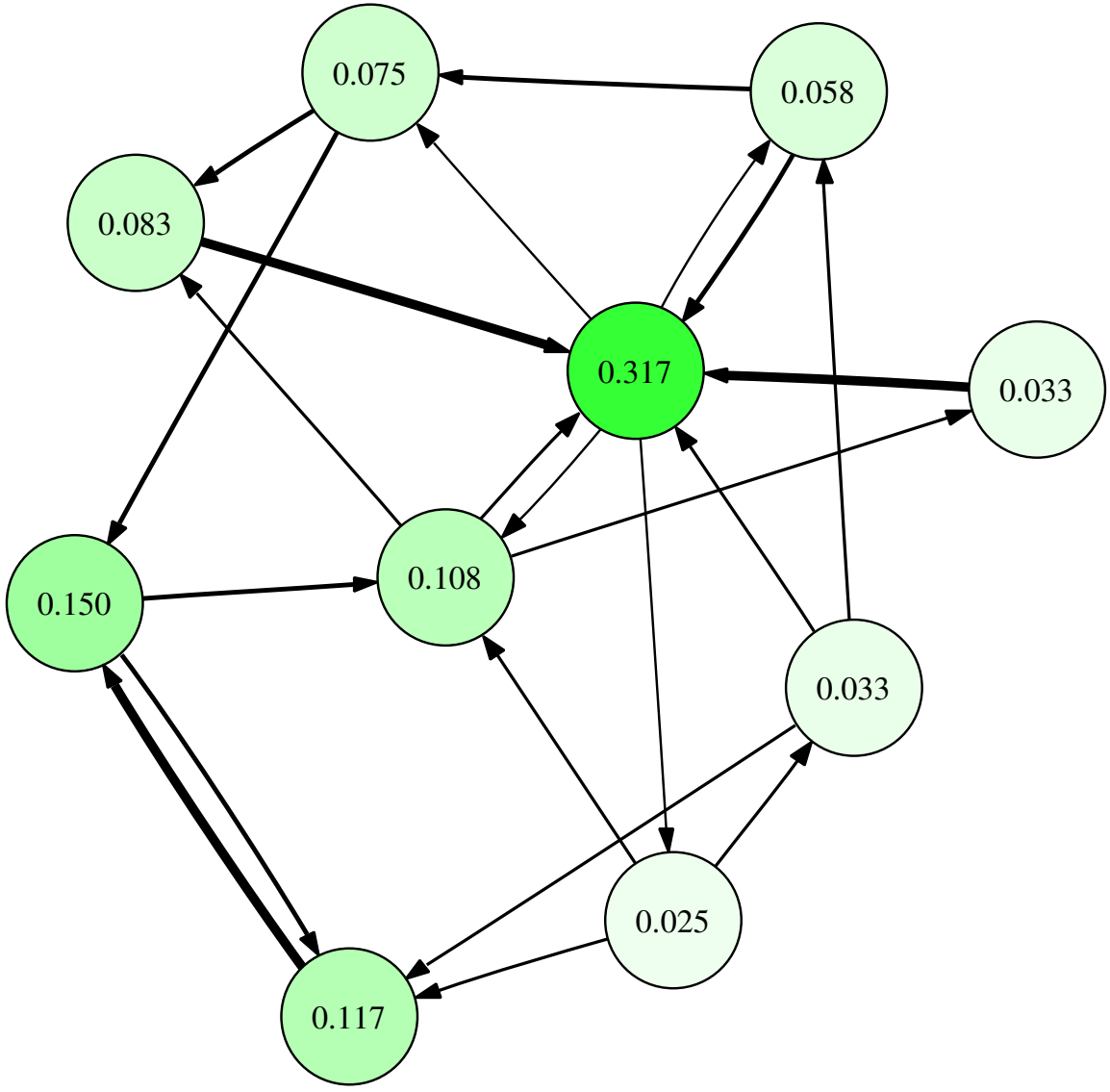
On peut montrer qu'il y a convergence du vecteur v vers une limite.

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

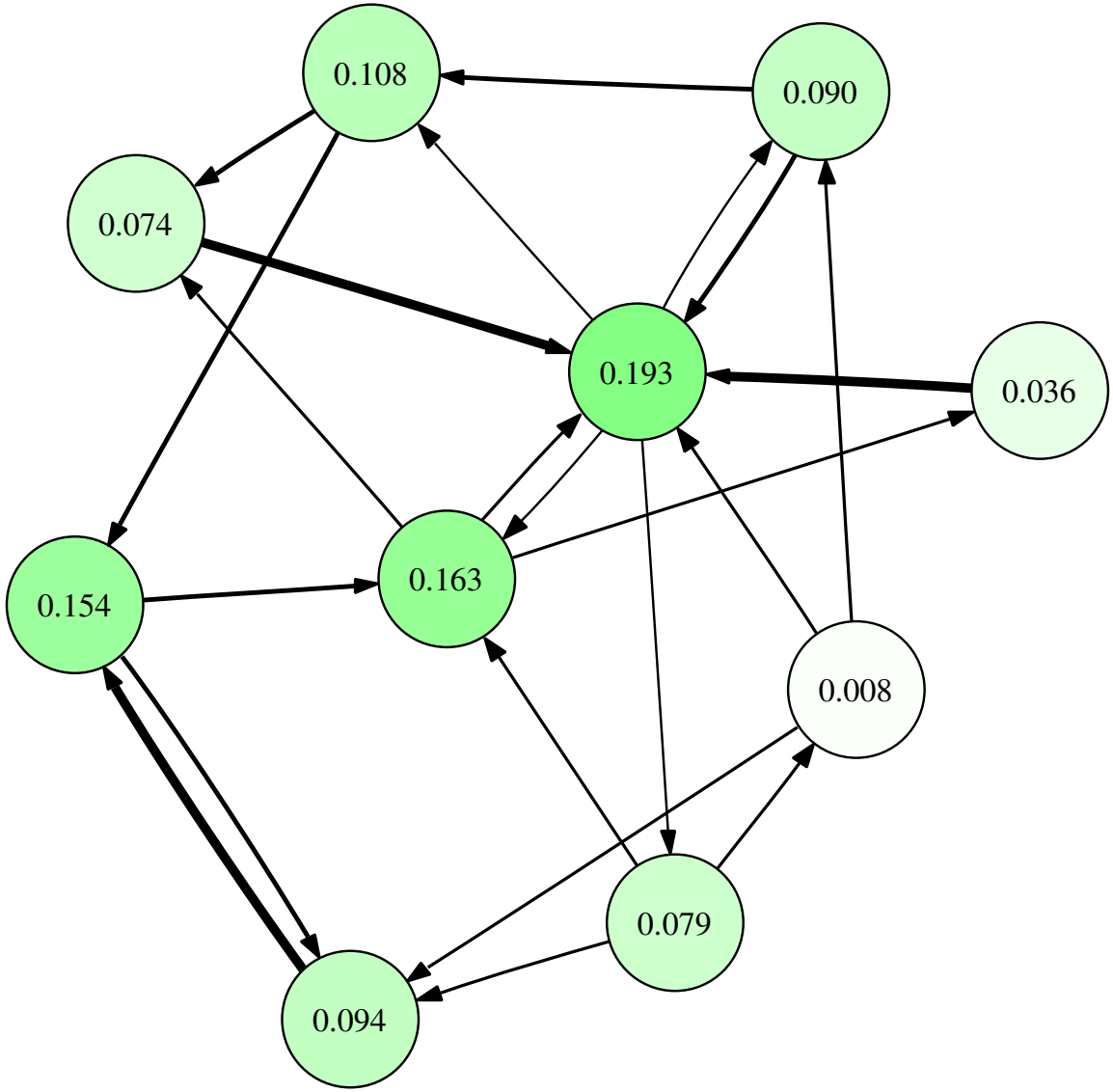
Quelques itérations PageRank



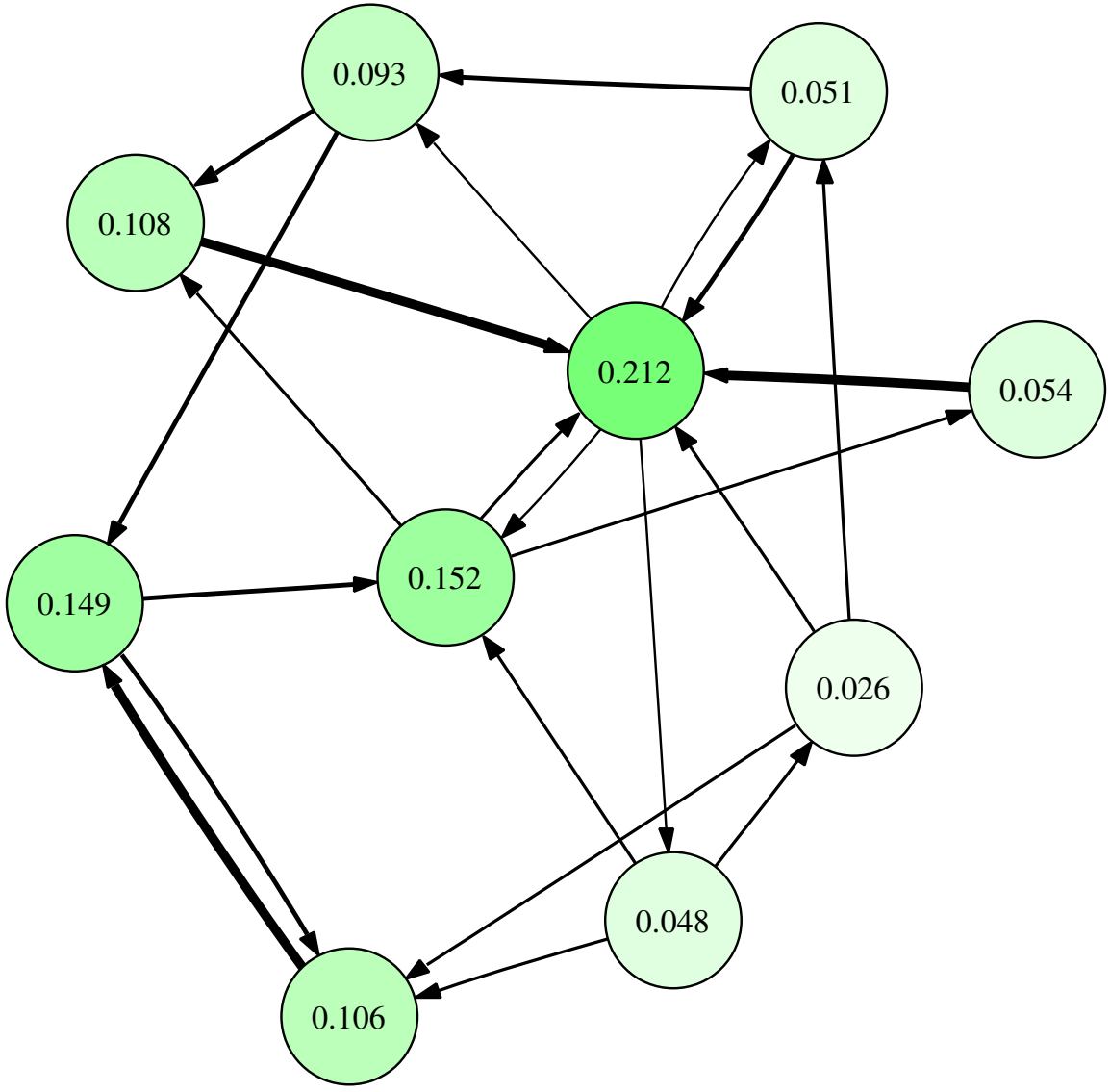
Quelques itérations PageRank



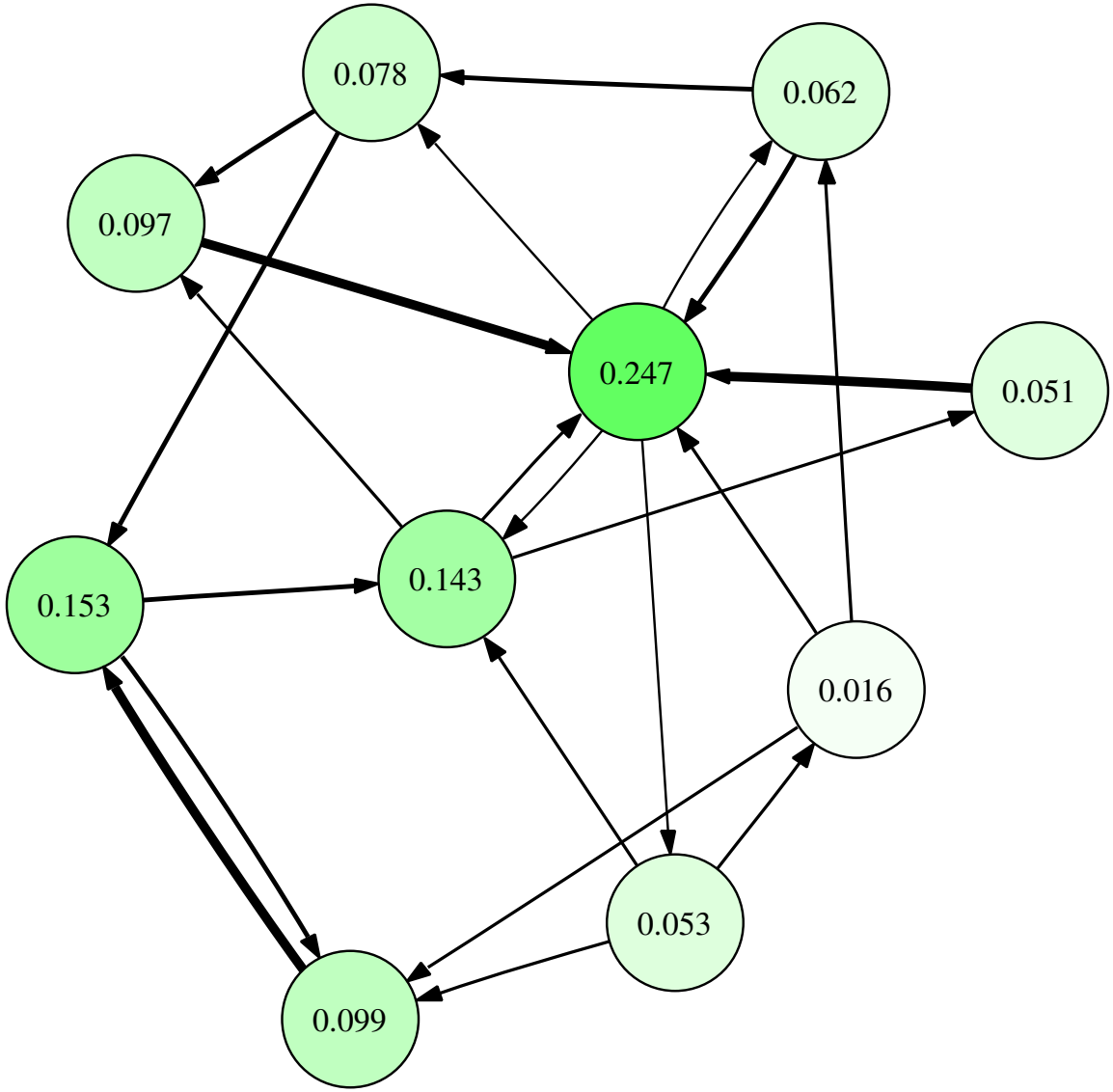
Quelques itérations PageRank



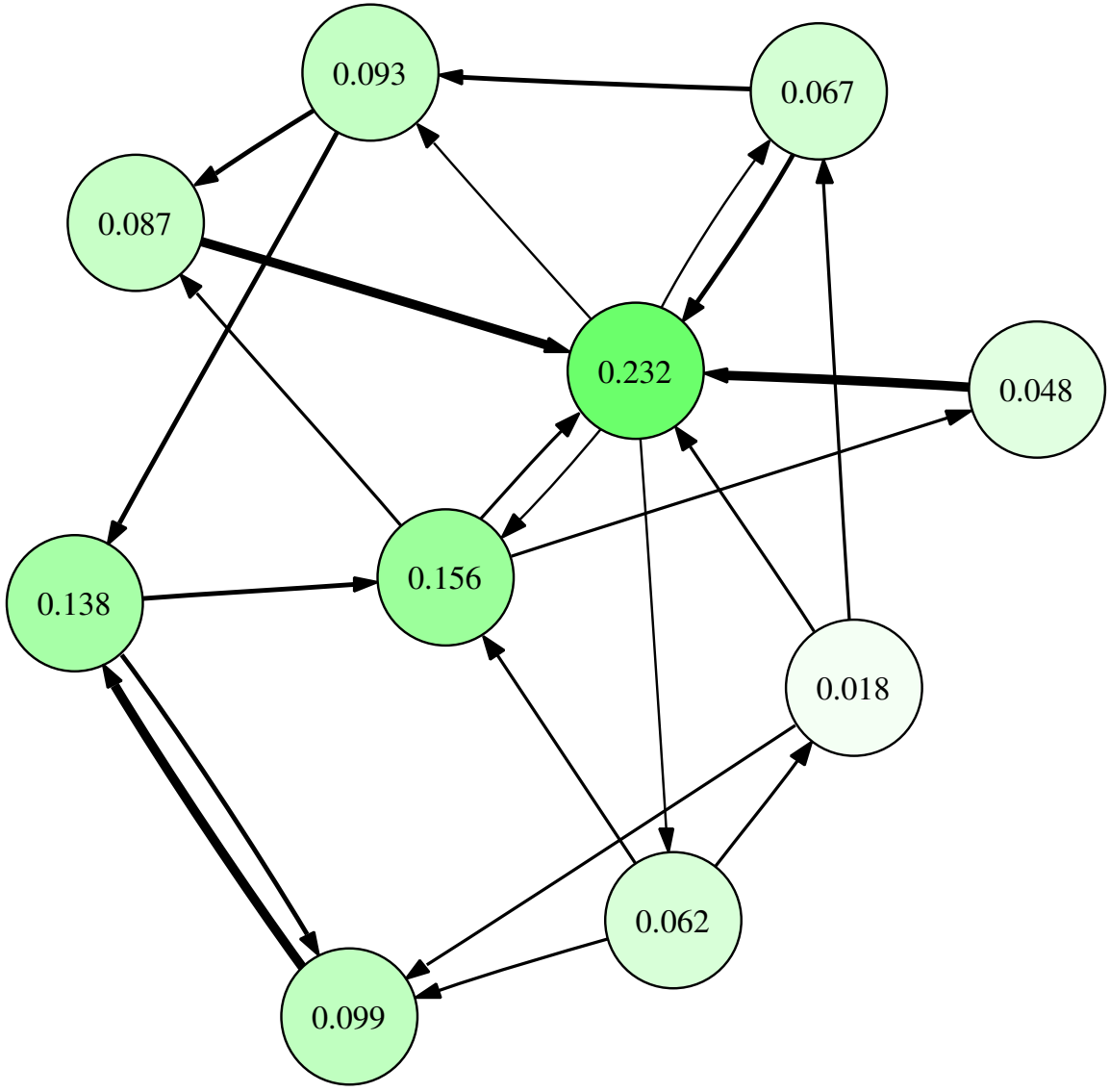
Quelques itérations PageRank



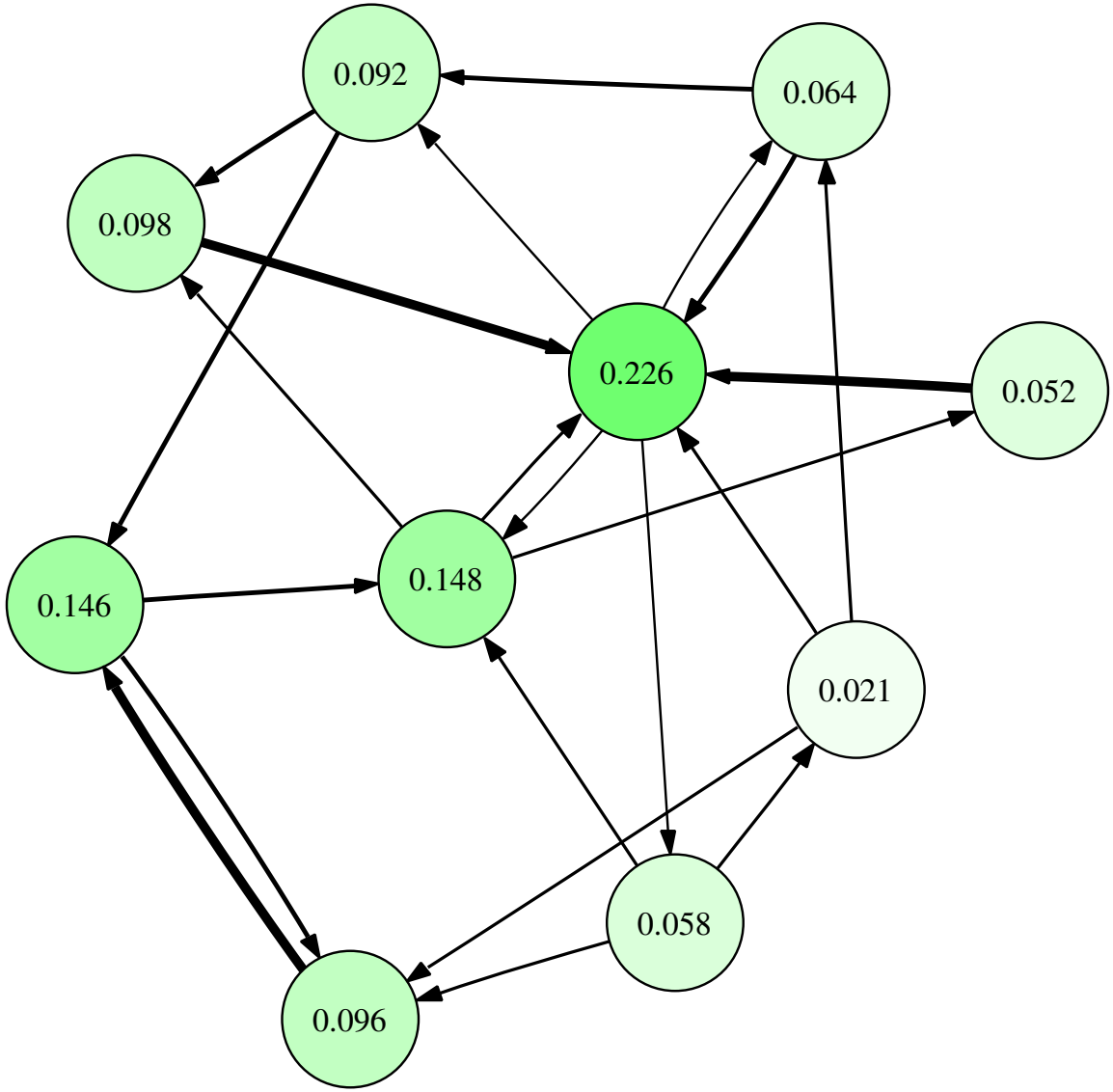
Quelques itérations PageRank



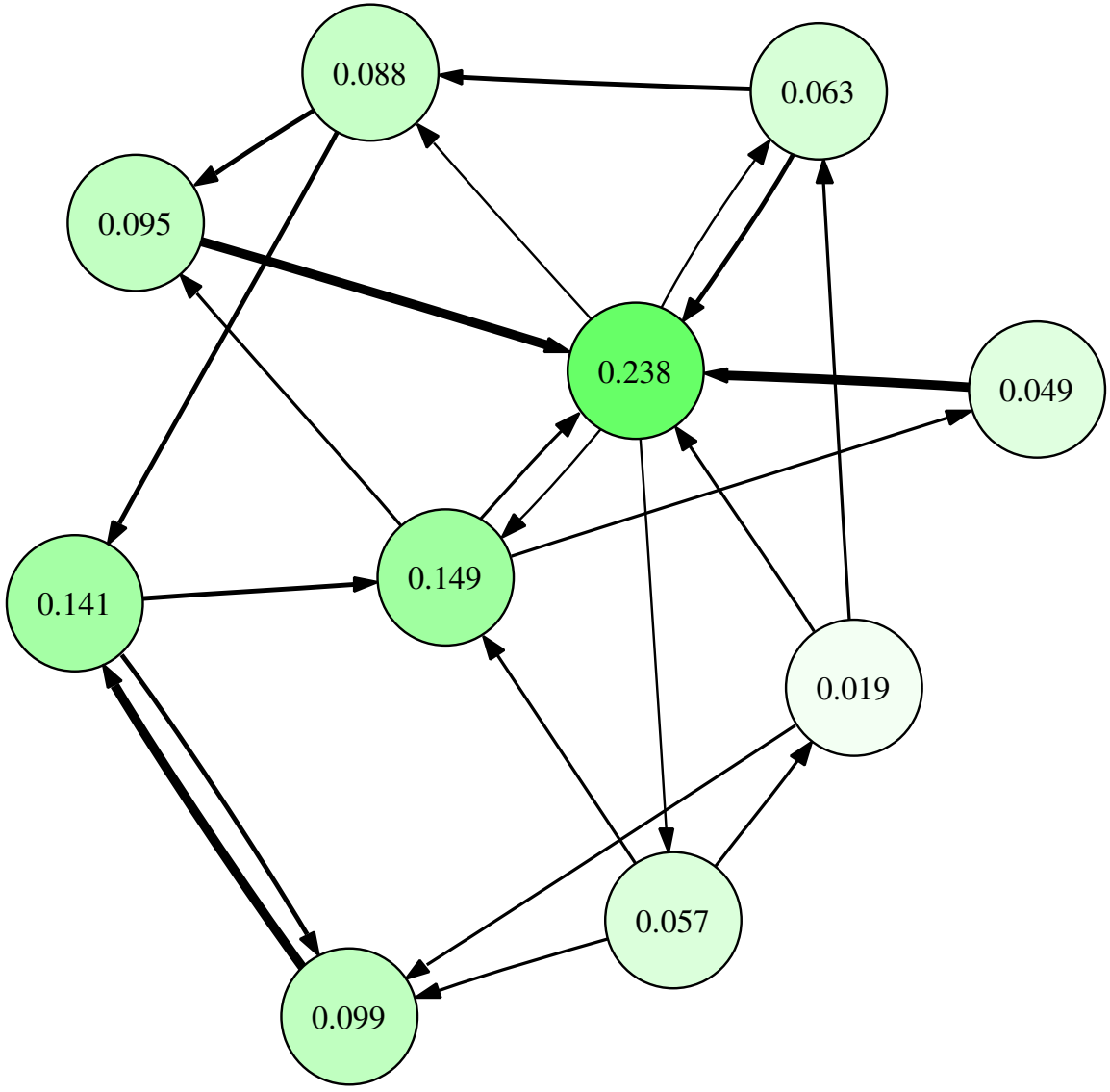
Quelques itérations PageRank



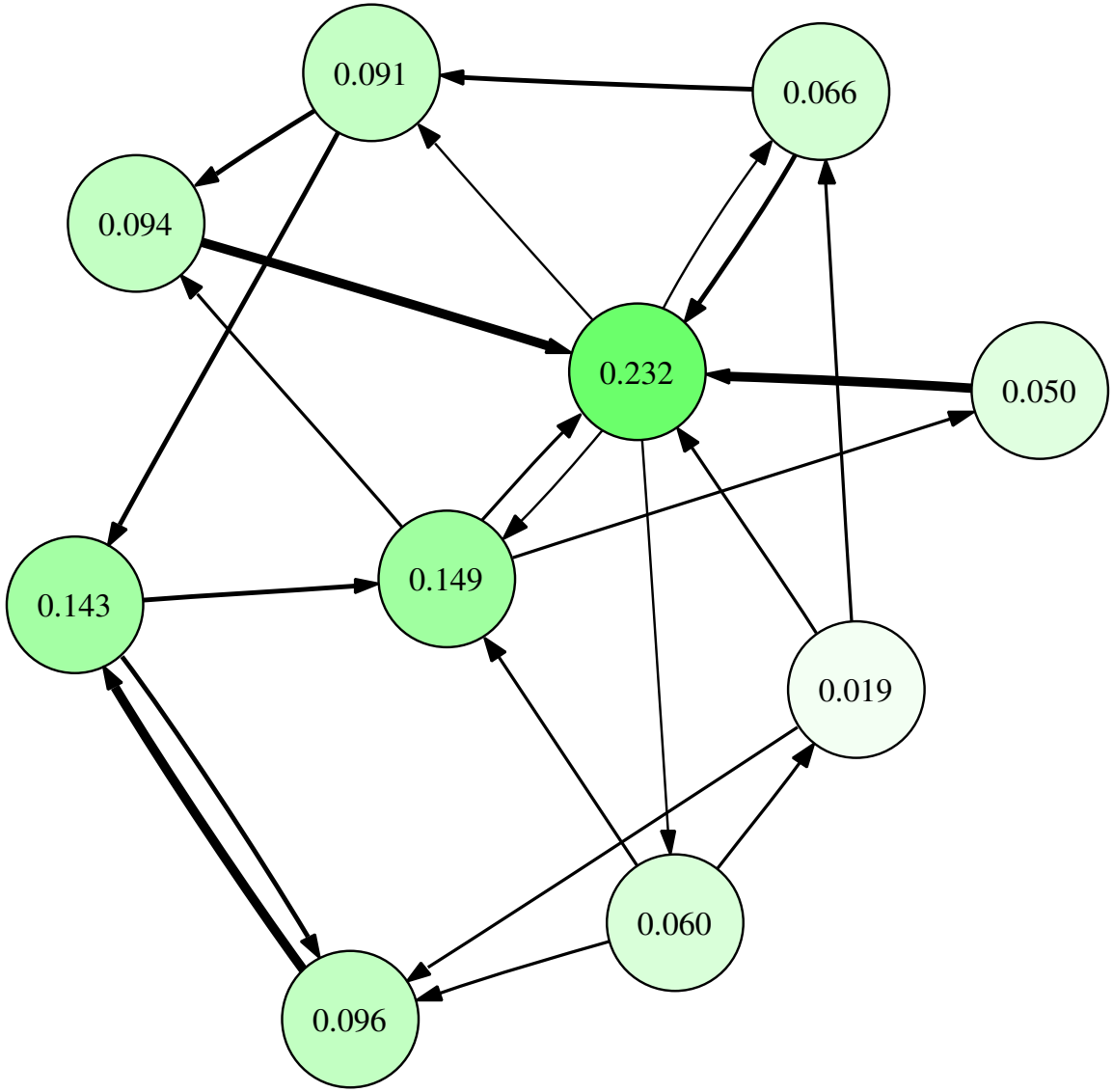
Quelques itérations PageRank



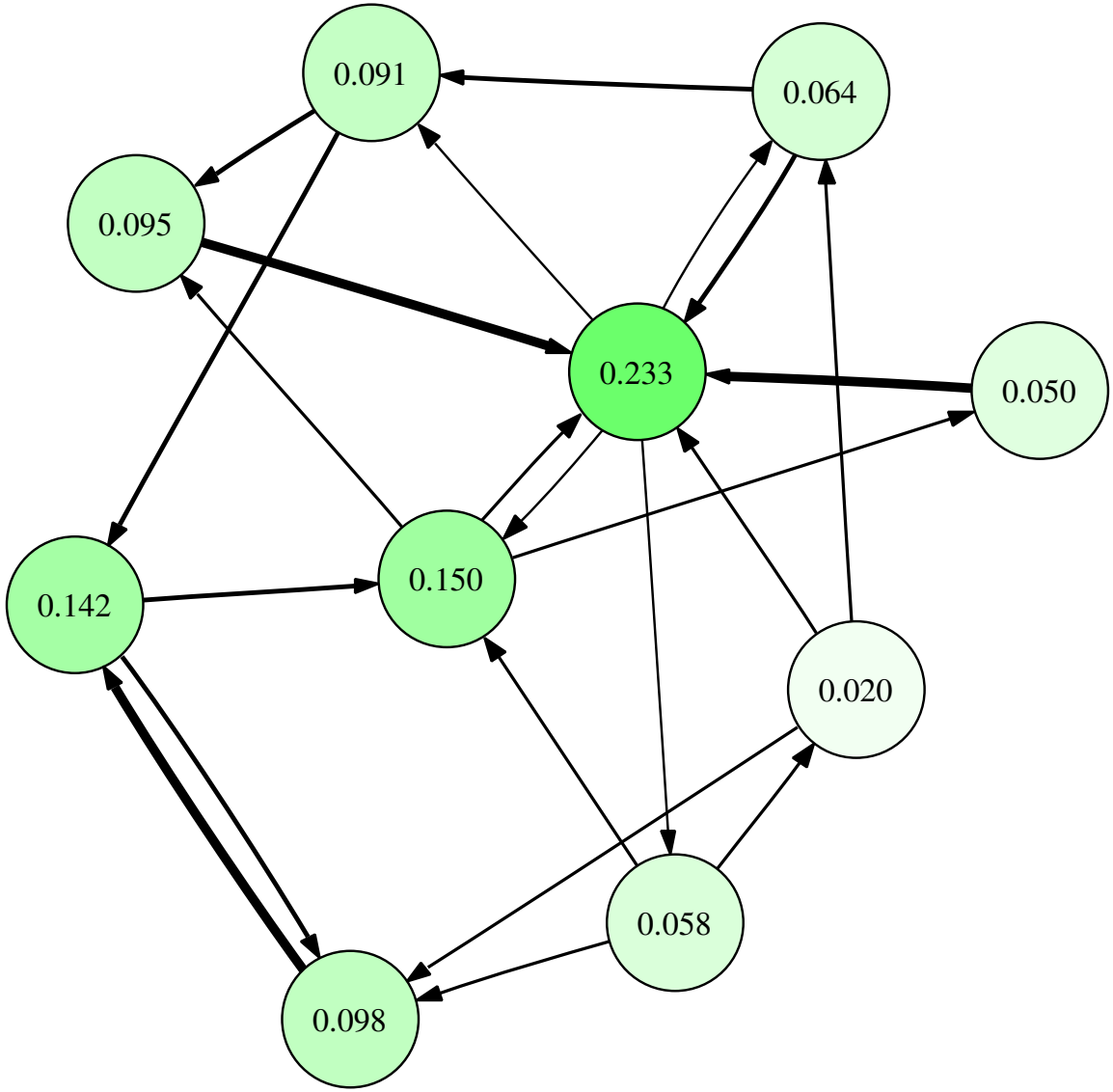
Quelques itérations PageRank



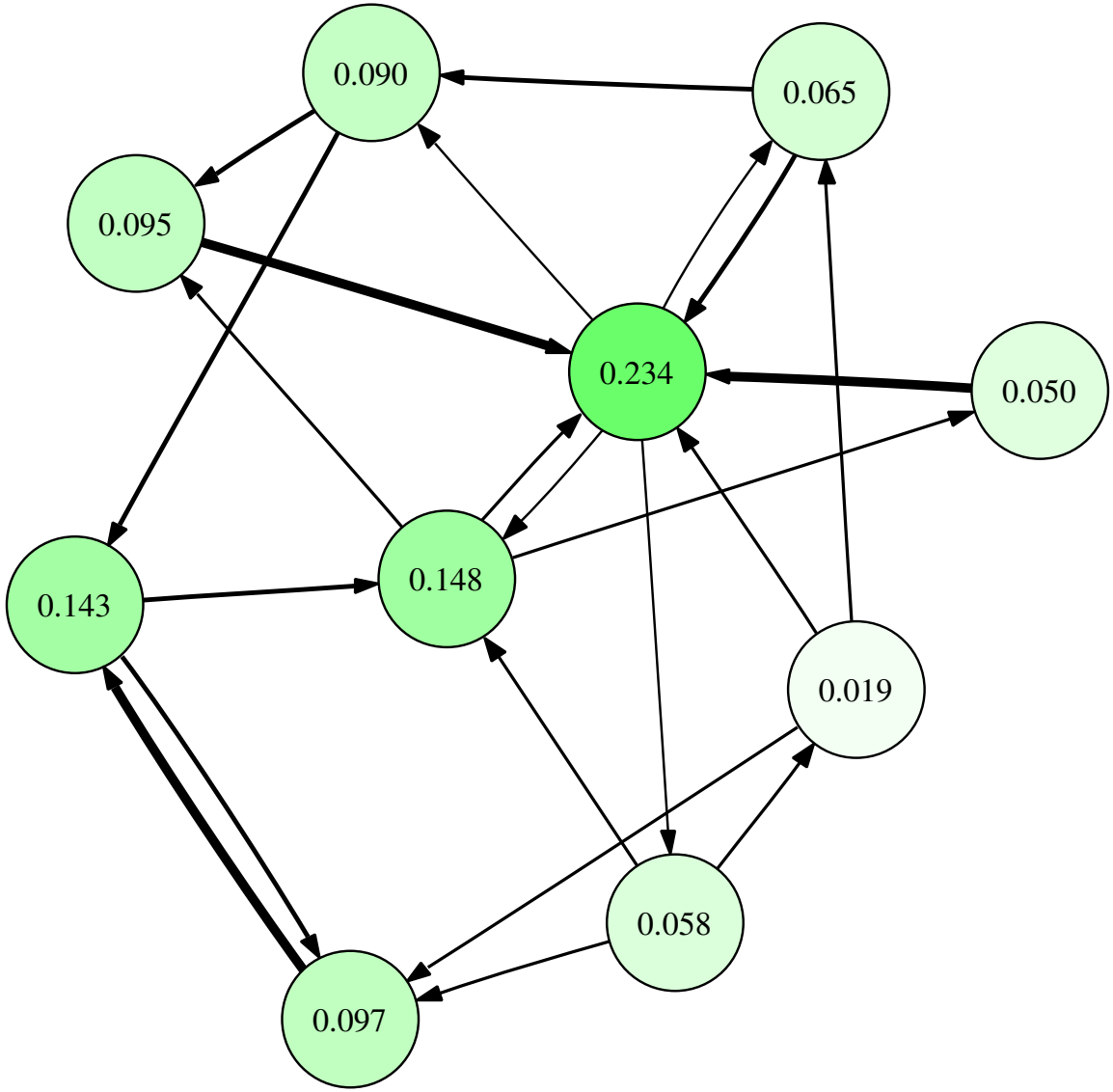
Quelques itérations PageRank



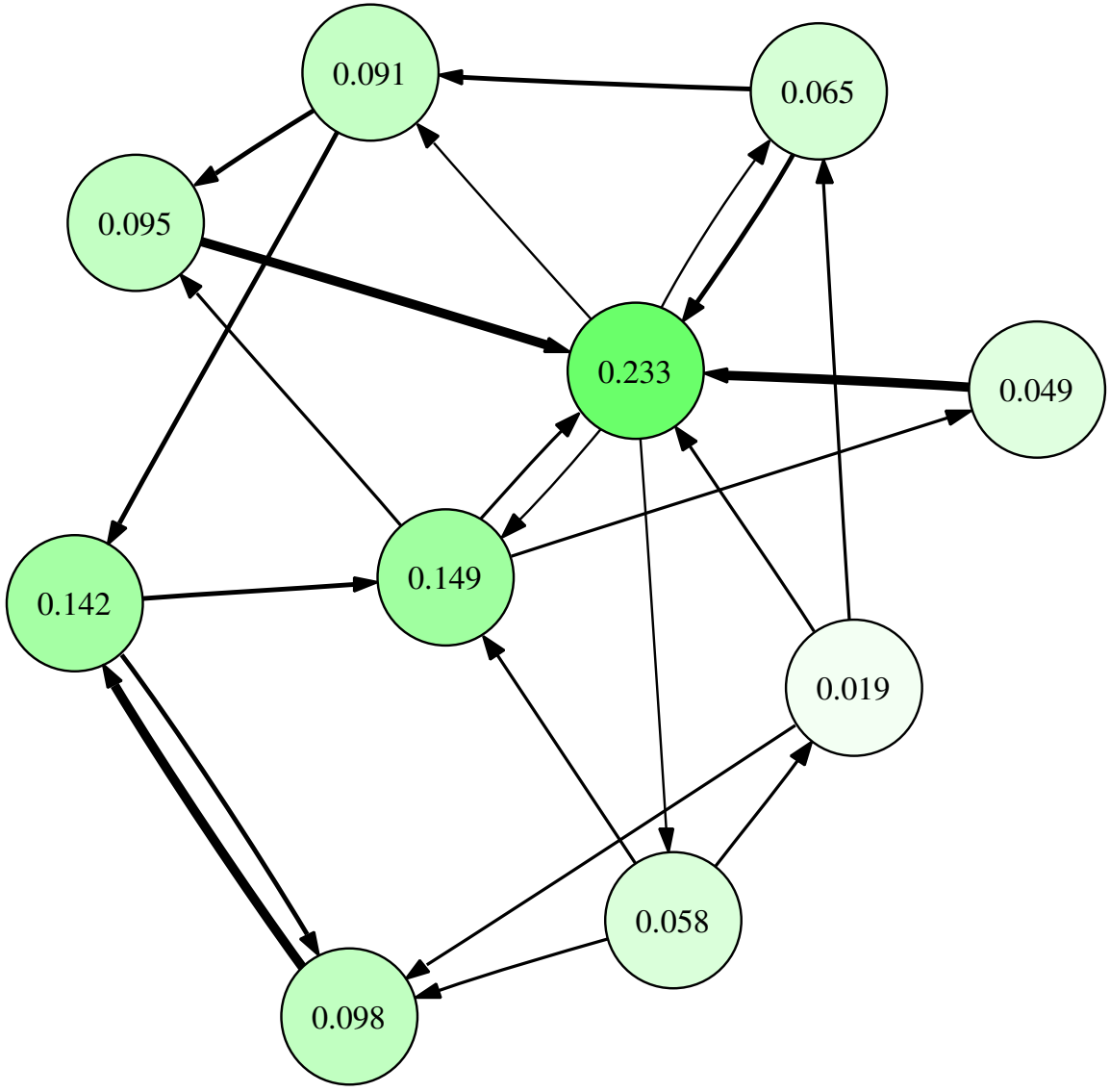
Quelques itérations PageRank



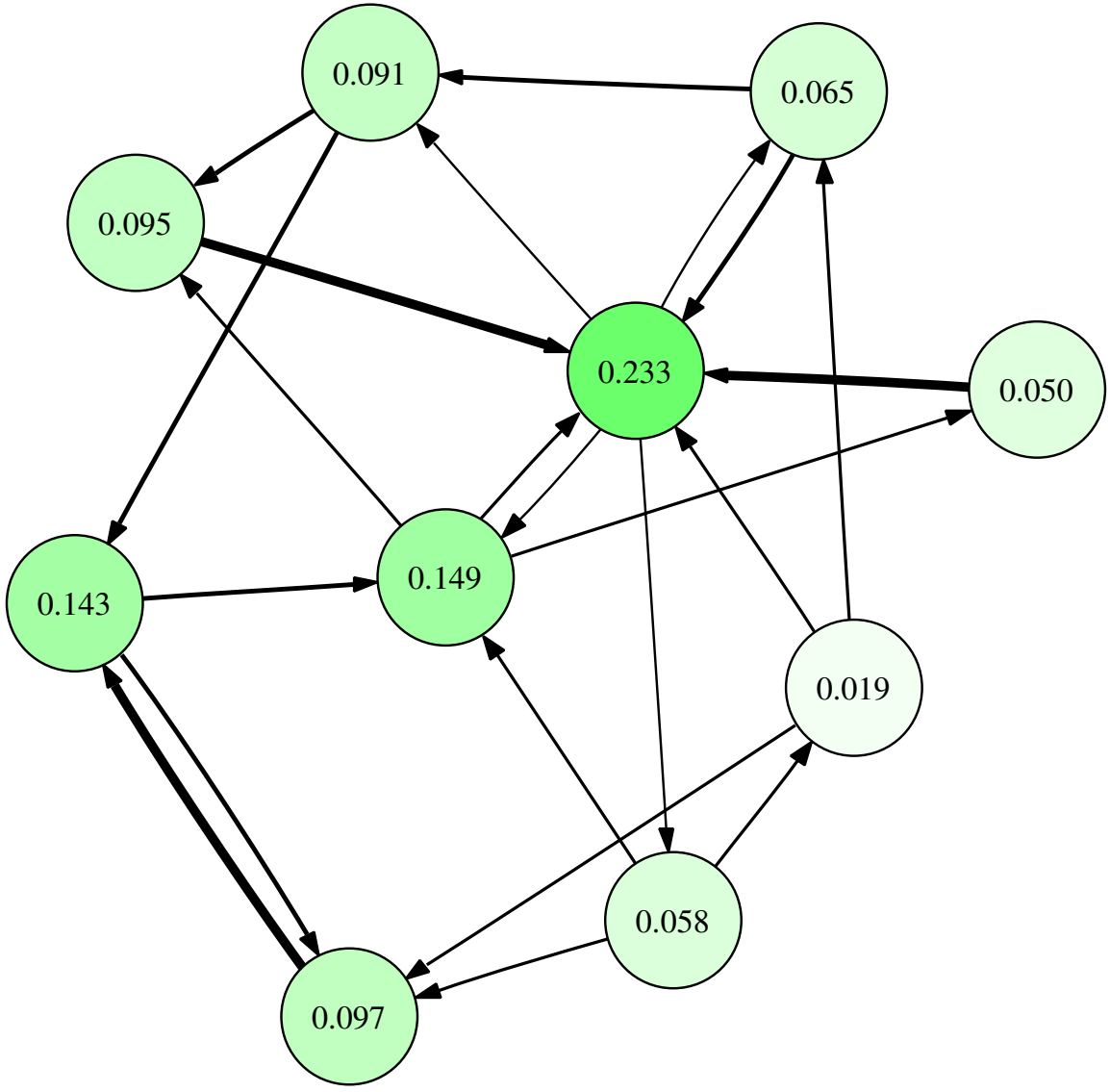
Quelques itérations PageRank



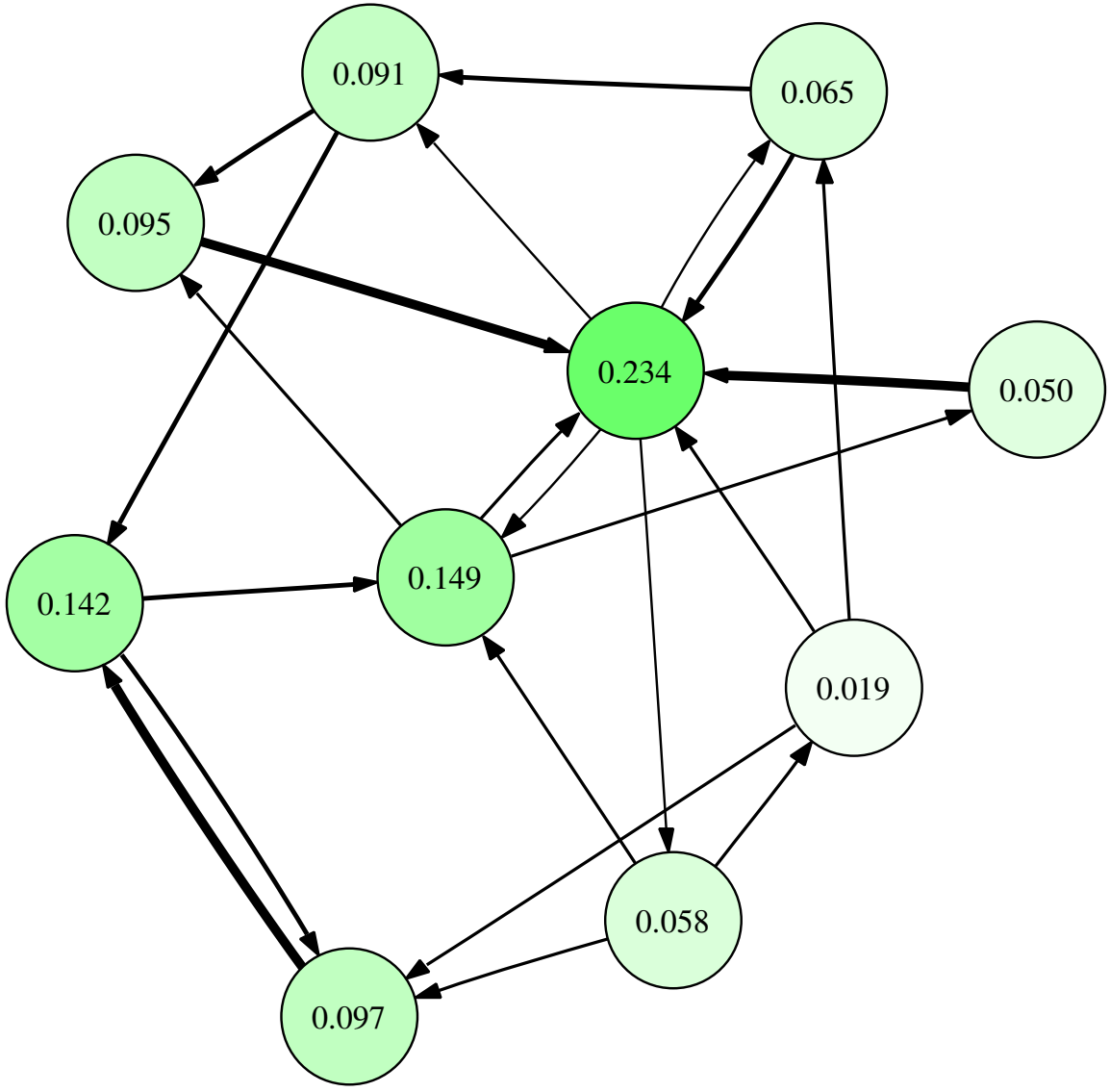
Quelques itérations PageRank



Quelques itérations PageRank



Quelques 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.

Centralité de Katz

- mesure le nombre de tous les nœuds connectés par un chemin, tout en pénalisant les contributions de nœuds éloignés

$$x_i = \sum_k \sum_j \alpha^k (A^k)_{ji}, \alpha \in [0,1]$$

- se réécrit en :

$$x_i = \alpha \sum_j^N A_{ij}(x_j + 1)$$

- centralité de vecteur propre avec $x_j + 1$ au lieu de x_j

- Le PageRank vérifie :

$$x_i = \alpha \sum_j^N A_{ij} \frac{x_j}{L(j)} + \frac{(1-\alpha)}{N}$$

- $L(j) = \sum_i A_{ij}$ nombre de voisins de j

Centralité d'intermédiarité (betweeness)

- On s'intéresse à la façon dont un nœud est important pour relier d'autres nœuds
- intermédiarité : Nombre de plus courts chemins passant par un sommet.

$$C(v) = \sum_{s \neq t \neq v \in V} \frac{s_{st}(v)}{s_{st}}$$

$s_{st}(v)$ est le nombre de plus courts chemins de s vers t
$s_{st}(v)$ est le nombre de plus courts chemins de s vers t, passant par v

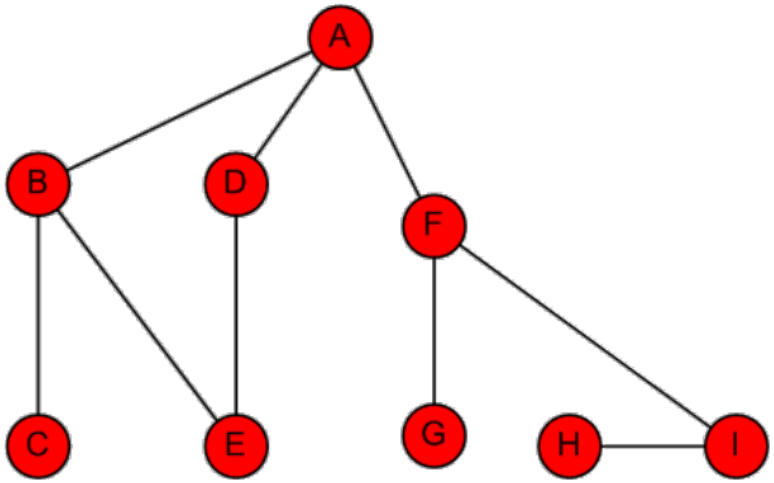
- Valeur maximale ? $\frac{s_{st}(v)}{s_{st}} = 1$

$$C(v) = \sum_{s \neq t \neq v \in V} 1 = 2\binom{n-1}{2} = (n-1)(n-2)$$

- On normalise :

$$C_b(v) = \frac{C(v)}{2\binom{n-1}{2}}$$

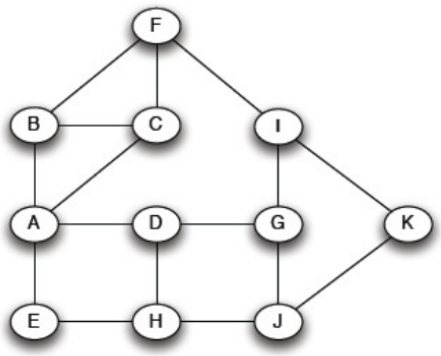
Intermédiarité : exemple

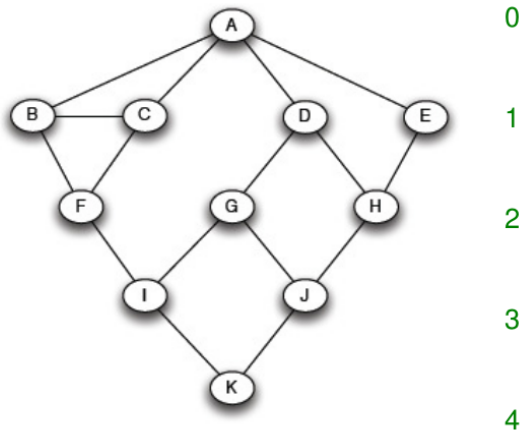


Node	Betweenness Centrality	Rank
A	16 + 1/2 + 1/2	**1**
B	7+5/2	**3**
C	0	**7**
D	5/2	**5**
E	1/2 + 1/2	**6**
F	15 + 2	**1**
G	0	**7**
H	0	**7**
I	7	**4**

Calcul de la centralité

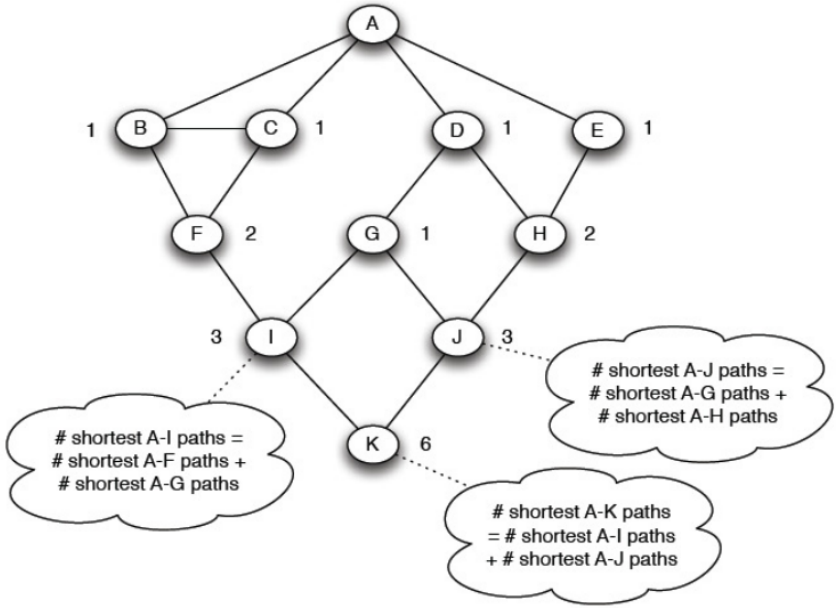
- On veut calculer la centralité des chemins commençant à A
- On fait un parcours en largeur à partir de A





Calcul de la centralité

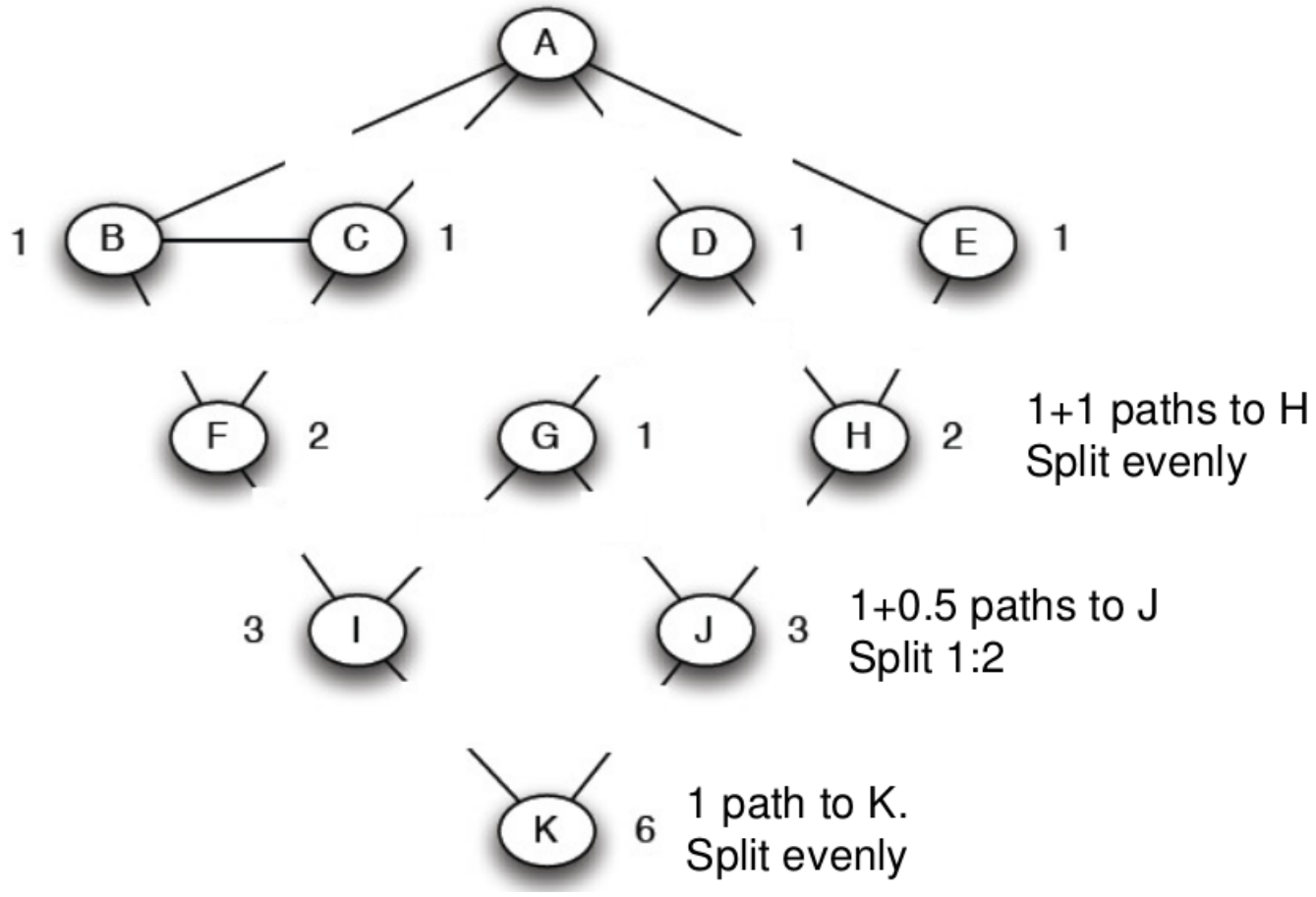
- la racine reçoit 1
- On donne à chaque nœud la somme des valeurs de ses parents
- (c'est le nombre de plus courts chemins arrivant à ce nœud)



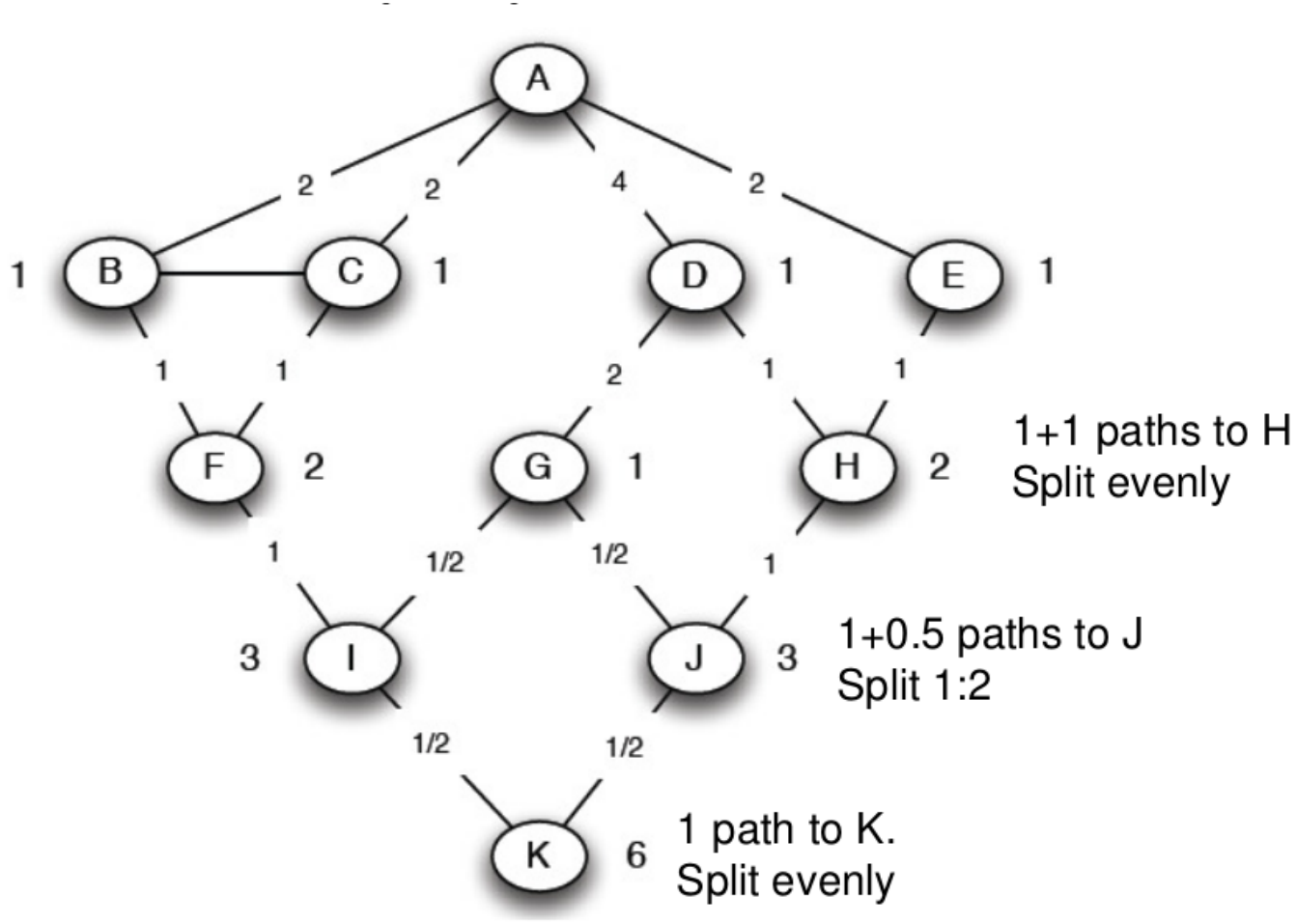
Calcul de la centralité

- Ensuite, en **remontant** l'arbre, on calcule la centralité
- chaque nœud reçoit 1 si c'est une feuille
- le "flot" d'un nœud vaut 1+$\sum$ "liens enfants"
- on le divise en fonction des "valeurs" des parents

Calcul


A
1
B
C
1
D
1
E
1
F
2
G
1
H
2
1+1 paths to H
Split evenly
3
I
J
3
1+0.5 paths to J
Split 1:2
K
6
1 path to K.
Split evenly

Calcul


A
2
2
4
2
1
B
C
1
D
1
E
1
1
1
2
1
1
F
2
G
1
H
2
1+1 paths to H
Split evenly
1
1/2
1/2
1
3
I
J
3
1+0.5 paths to J
Split 1:2
1/2
1/2
K
6
1 path to K.
Split evenly

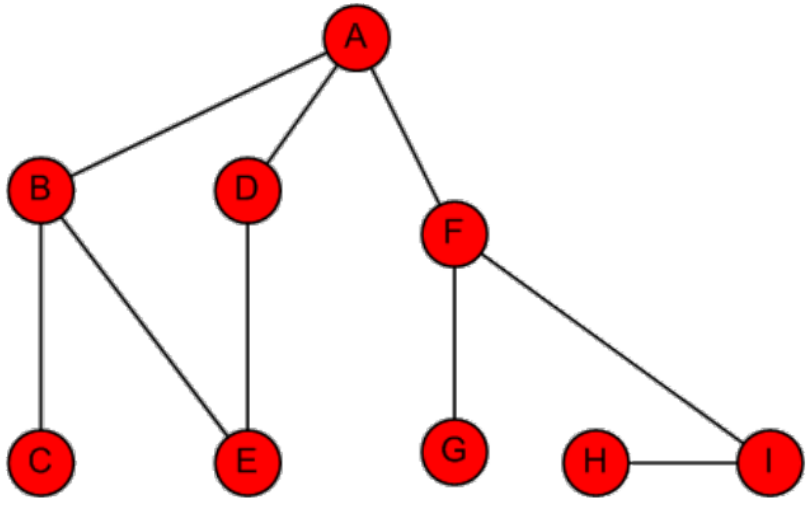
Centralité de proximité

- Repose sur l'intuition que les nœuds centraux doivent être ceux qui peuvent atteindre rapidement d'autres nœuds
- La moyenne des plus courts chemins de ces nœuds vers les autres doit être plus faible que les autres :

$$C(x) = \frac{1}{\bar{l_v}}$$

avec $\bar{l_v} = \frac{1}{n-1} \sum_{u \neq v} d(u, v)$

Centralité de proximité : exemple



Node	A	B	C	D	E	F	G	H	I	D_Avg	Closeness Centrality	Rank
A	0	1	2	1	2	1	2	3	2	1.750	0.571	**1**
B	1	0	1	2	1	2	3	4	3	2.125	0.471	**3**
C	2	1	0	3	2	3	4	5	4	3.000	0.333	**8**
D	1	2	3	0	1	2	3	4	3	2.375	0.421	**4**
E	2	1	2	1	0	3	4	5	4	2.750	0.364	**7**
F	1	2	3	2	3	0	1	2	1	1.875	0.533	**2**
G	2	3	4	3	4	1	0	3	2	2.750	0.364	**7**
H	3	4	5	4	5	2	3	0	1	3.375	0.296	**9**
I	2	3	4	3	4	1	2	1	0	2.500	0.400	**5**

Synthèse des centralités

	Faible degré	**Faible proximité**	**Faible intermédiarité**
Grand degré		Nœud dans une communauté loin du reste du réseau	Les connexions du nœud sont redondantes, les communications contournent le nœud
Grande proximité	Nœud-clé, connecté à des acteurs importants		Il y a beaucoup de chemins dans le réseau, le nœud est proche de beaucoup d'autres, mais de nombreux autres sont dans son cas
Grande intermédiarité	Les quelques voisins du nœud sont cruciaux pour la circulation dans le graphe	Très rare, le nœud est LE point de passage de peu d'acteurs vers beaucoup d'autres	

Centralité de groupe

- Pour le moment, on n'a défini les centralités que pour un nœud
- On peut généraliser à un groupe de nœuds
- On remplace pour cela le groupe par un "super nœud", sans tenir compte de la structure interne du groupe
- Avec S le groupe, et $V - S$ l'ensemble des autres nœuds :

- centralité de degrés :

$$C^g(S) = |\{v \in V - S \text{ t.q. } v \text{est connecté à } u \in S\}|$$

- intermédiarité :

$$C^g(S) = \sum_{s \neq t, s \notin S, t \notin S} \frac{s_{st}(S)}{s_{st}}$$

- proximité :

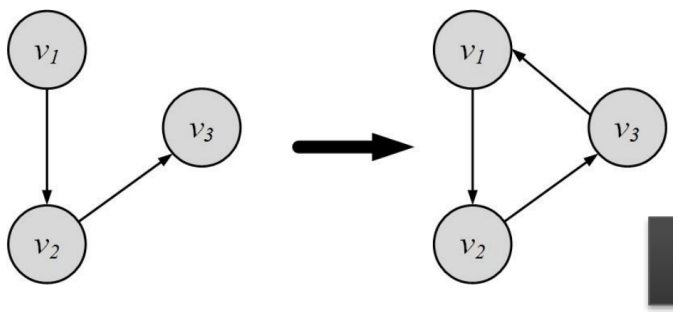
$$C^g(S) = \frac{1}{l_S^g}$$

l_S^g est la distance moyenne des non-membres envers le groupe (ou min, ou max)

Transitivité

- Un ami de mes amis est mon ami
- la transitivité densifie le graphe, le rapproche d'un graphe complet
- on mesure à quel point les graphes sont proches de graphes complet en mesurant la transisivité

- On mesure des coefficients de clustering, globaux et locaux

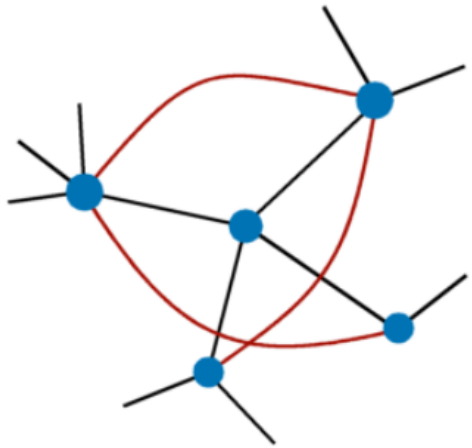


Coefficient de clustering

- Global

$$C = \frac{3 \times |\text{triangles}|}{|\text{paires de voisins distincts}|} = \frac{3 \times |\text{triangles}|}{\sum_{i \in V} \binom{d_i}{2}}$$

- Local



$c(i) = \frac{2*|(x,y) \in E, x,y \in N(i)|}{k_i(k_i-1)}$ (ou 0 si $d(i) < 2$))

Similarité et équivalence

- On cherche à quel point deux nœuds du graphe sont structurellement équivalents
- on examine les nœuds qui sont communs à leurs voisinages
- la taille de ce voisinage partagé définit leur similarité

- Mesure brute :

$$\sigma(u,v) = |N(u) \cap N(v)|$$

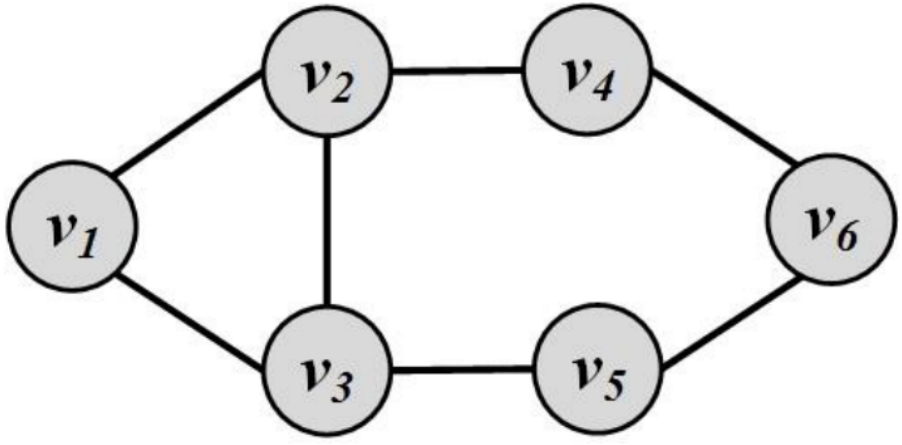
- Normalisation : indice de Jaccard

$$\sigma(u,v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

- Normalisation : similarité cosinus

$$\sigma(u,v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)||N(v)|}}$$

Exemple



$$\sigma_{Jaccard}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cup \{v_3, v_6\}|}{|\{v_1, v_3, v_4, v_6\}|} = 0.25$$

$$\sigma_{cosinus}(v_2, v_5) = \frac{|\{v_1, v_3, v_4\} \cup \{v_3, v_6\}|}{\sqrt{|\{v_1, v_3, v_4\}||\{v_3, v_6\}|}} = 0.4$$

Significativité

- On compare la similarité calculée avec la valeur espérée si les nœuds choisissaient leurs voisins au hasard
- Pour des nœuds u et v de degré d_u et d_v, cette valeur est : $\frac{d_u d_v}{n}$
  - v a une chance d_u/n de devenir le voisin de u
  - v choisit d_v sommets
- on peut réécrire ça :

$$\sigma_{sign} = \sum_k A_{ik} A_{kj} - \frac{d_i d_j}{n}$$

(proche du coefficient de Pearson)

Équivalence régulière

- On peut aussi chercher une similarité où u et v sont similaires si leurs voisins sont similaires (récursivité)

Modélisation

Introduction : pourquoi des modèles ?

- Pour analyser et comprendre un graphe, les modèles sont un puissant outil
- On cherche à générer des graphes qui sont similaires à des graphes réels

- On peut alors simuler des graphes (coût)
- On peut proposer des explications mathématiquement validées à des phénomènes
- On peut mener des expériences contrôlées (en l'absence de graphes mesurés, par exemple)

Quelles sont les propriétés que l'on veut modéliser ?

Propriétés des réseaux réels

- distribution de degré fortement hétérogène
- faible densité
- fort clustering
- faible distance moyenne
- composante géante
- présence de communautés

propriétés différentes de celles des graphes aléatoires simples

Loi de puissance

- Quand la fréquence d'un événement change avec la puissance d'un attribut, la fréquence suit une **loi de puissance**

$$p_d = ad^{-b}$$

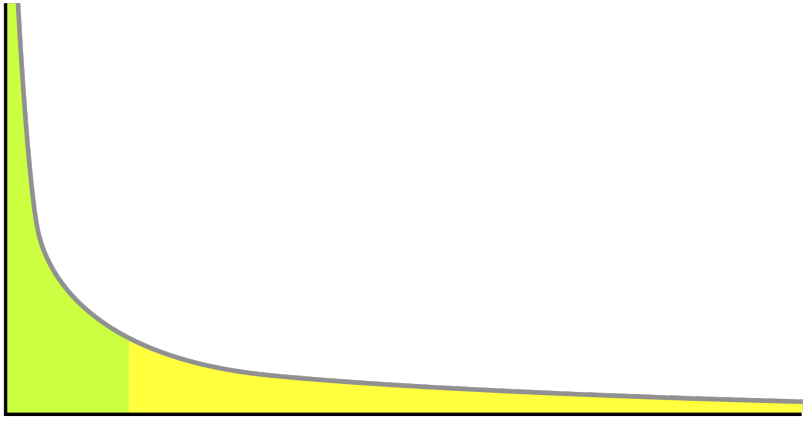
- où :
  - p_d fraction d'utilisateurs de degré d
  - b est l'exposant, typiquement entre 2 et 3
  - d est le degré
  - a est une constante de proportionnalité

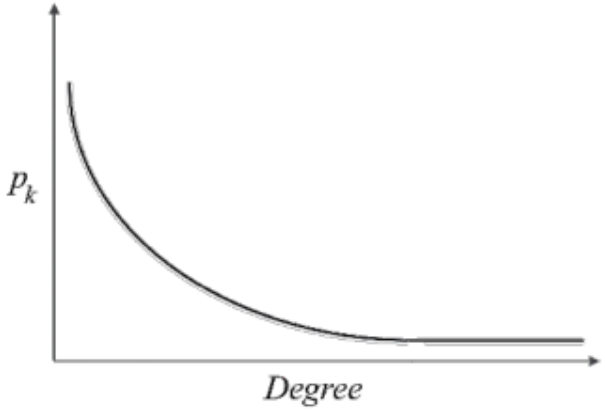
- on a :

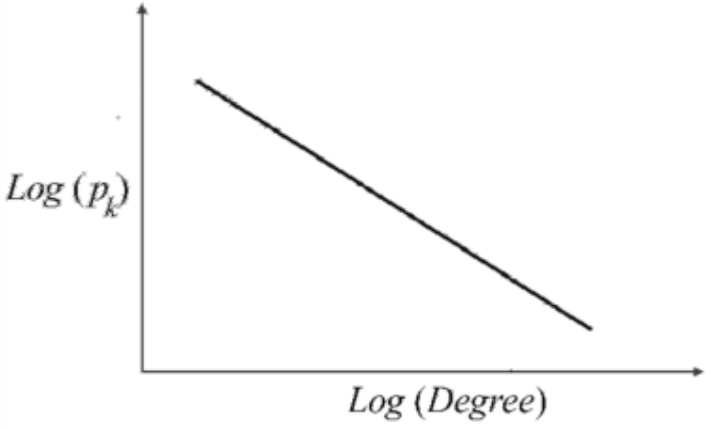
$$ln(p_d) = -b\ ln(d) + ln(a)$$

Loi de puissance : exemple

- on retrouve les lois de puissance quand la quantité mesurée ressemble à de la popularité
- avec une loi de puissance :
  - il y a beaucoup de faibles occurrences
  - les grandes occurrences sont rares
  - ”long tail”





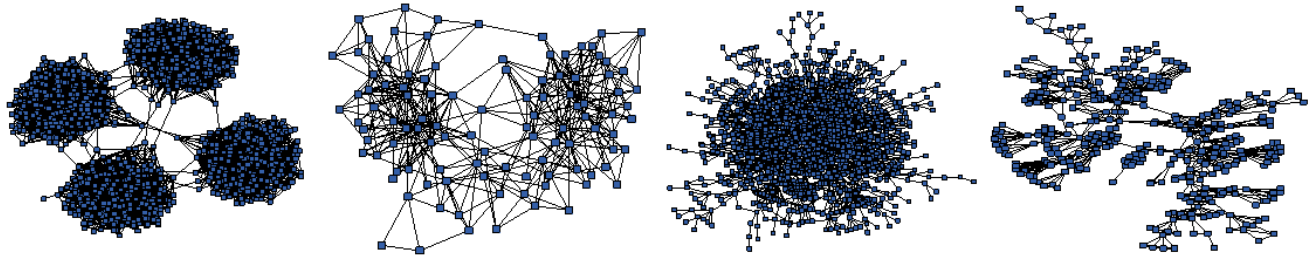


Caractéristiques des lois de puissance

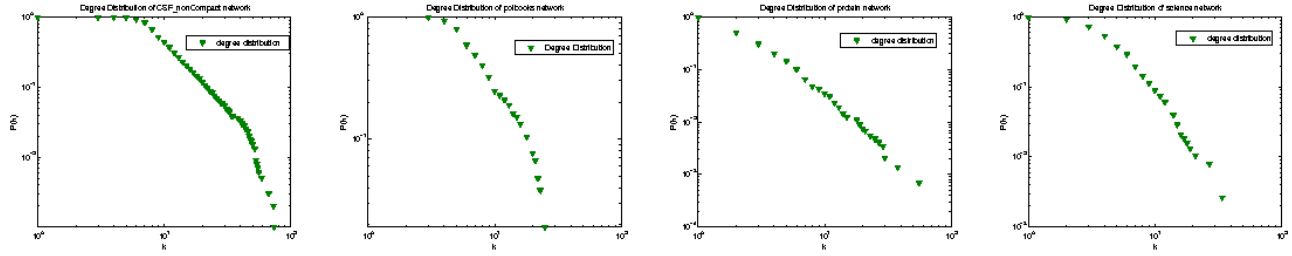
- On parle d'invariance d'échelle (en anglais : scale-free networks)
- Si le degré est multiplié par α, on a :

$$p_{\alpha d} = a(\alpha d)^{-b} = \alpha^{-b} \cdot p_d$$

- les moments sont particuliers :
  - moyenne définie si exposant > 2
  - variance définie si exposant > 3
  - (attention aux statistiques)



(b) CSF-NonCompact (c) Polbooks (d) Protein (e) Science



(b) CSF-NonCompact (c) Polbooks (d) Protein (e) Science

Quelques domaines d'apparitions

Les lois de puissances se retrouvent :

- dans les réseaux d'appels, où la fraction des numéros qui reçoivent k appels par jour est à peu près proportionnelle à $1/k^2$
- en librairies : la fraction de livres qui est achetée par k personnes est à peu près proportionnelle à $1/k^3$
- en bibliométrie : la fraction d'article de recherche qui reçoivent k citations est proportionnelle à $1/k^3$
- en réseaux sociaux (Twitter, Facebook) : la fraction d'utilisateurs qui ont un degré entrant de k est à peu près proportionnelle à $1/k^2$

Coefficient de clustering

- les amitiés dans les réseaux sociaux sont très transitives : les amis d'une personne sont souvent amis entre eux
- cela forme des triades, avec un coefficient de clustering local élevé (densité)

	Network	Type	n	m	C
Social	Film actors	Undirected	449 913	25 516 482	0.20
	Company directors	Undirected	7 673	55 392	0.59
	Math coauthorship	Undirected	253 339	496 489	0.15
	Physics coauthorship	Undirected	52 909	245 300	0.45
	Biology coauthorship	Undirected	1 520 251	11 803 064	0.088
	Telephone call graph	Undirected	47 000 000	80 000 000	
	Email messages	Directed	59 812	86 300	
	Email address books	Directed	16 881	57 029	0.17
	Student dating	Undirected	573	477	0.005
	Sexual contacts	Undirected	2 810		
Information	WWW nd.edu	Directed	269 504	1 497 135	0.11
	WWW AltaVista	Directed	203 549 046	1 466 000 000	
	Citation network	Directed	783 339	6 716 198	
	Roget's Thesaurus	Directed	1 022	5 103	0.13
	Word co-occurrence	Undirected	460 902	16 100 000	
Technological	Internet	Undirected	10 697	31 992	0.035
	Power grid	Undirected	4 941	6 594	0.10
	Train routes	Undirected	587	19 603	
	Software packages	Directed	1 439	1 723	0.070
	Software classes	Directed	1 376	2 213	0.033
	Electronic circuits	Undirected	24 097	53 248	0.010
	Peer-to-peer network	Undirected	880	1 296	0.012
Biological	Metabolic network	Undirected	765	3 686	0.090
	Protein interactions	Undirected	2 115	2 240	0.072
	Marine food web	Directed	134	598	0.16
	Freshwater food web	Directed	92	997	0.20
	Neural network	Directed	307	2 359	0.18

Source: M. E. J Newman

Longueur moyenne des chemins

- Dans les réseaux sociaux, la longueur moyenne des chemins est en moyenne faible
- Expérience de Milgram (lettres), avec en moyenne 6 liens/5 intermédiaires

	Network	Type	n	m	ℓ
Social	Film actors	Undirected	449 913	25 516 482	3.48
	Company directors	Undirected	7 673	55 392	4.60
	Math coauthorship	Undirected	253 339	496 489	7.57
	Physics coauthorship	Undirected	52 909	245 300	6.19
	Biology coauthorship	Undirected	1 520 251	11 803 064	4.92
	Telephone call graph	Undirected	47 000 000	80 000 000	
	Email messages	Directed	59 812	86 300	4.95
	Email address books	Directed	16 881	57 029	5.22
	Student dating	Undirected	573	477	16.01
	Sexual contacts	Undirected	2 810		
Information	WWW nd.edu	Directed	269 504	1 497 135	11.27
	WWW AltaVista	Directed	203 549 046	1 466 000 000	16.18
	Citation network	Directed	783 339	6 716 198	
	Roget's Thesaurus	Directed	1 022	5 103	4.87
	Word co-occurrence	Undirected	460 902	16 100 000	
Technological	Internet	Undirected	10 697	31 992	3.31
	Power grid	Undirected	4 941	6 594	18.99
	Train routes	Undirected	587	19 603	2.16
	Software packages	Directed	1 439	1 723	2.42
	Software classes	Directed	1 376	2 213	5.40
	Electronic circuits	Undirected	24 097	53 248	11.05
	Peer-to-peer network	Undirected	880	1 296	4.28
Biological	Metabolic network	Undirected	765	3 686	2.56
	Protein interactions	Undirected	2 115	2 240	6.80
	Marine food web	Directed	134	598	2.05
	Freshwater food web	Directed	92	997	1.90
	Neural network	Directed	307	2 359	3.97

Source: M. E. J Newman

Autres propriétés

- Paradoxe des amis : en moyenne, vos amis ont davantage d'amis que vous
  - les nœuds de fort degré apparaissent dans de nombreuses moyennes
  - vrai pour 98% des utilisateurs Twitter (2013)
- Structure cœur-périphérie
  - cœur dense
  - périphérie avec des nœuds connectés au cœur, pas entre eux

Graphes aléatoires

Modèle $G(n, p)$

- On considère un graphe avec un nombre n de nœuds
- Il peut y avoir $\binom{n}{2}$ arêtes, elles ont une probabilité p d'exister
- On appelle ce graphe un graphe aléatoire $G(n, p)$

Modèle $G(n, m)$, Erdos-Rényi

- On suppose n et m fixés
- Il faut déterminer quelles arêtes sont sélectionnées parmi les possibles
- Soit Ω l'ensemble des graphes avec n sommets et m arêtes.
- Il y a $|\Omega| = \binom{\binom{n}{2}}{m}$ graphes possibles
- On génère un graphe parmi ceux-là

Comparaisons

- Quand n est grand, les deux modèles sont proches :
  - le nombre d'arêtes dans $G(n,p)$ est $\binom{n}{2}p$
  - si l'on fixe $m = \binom{n}{2}p$, on a les mêmes résultats avec $G(n,m)$
- Différences
  - Le modèle $G(n,m)$ a un nombre fixe d'arêtes
  - Le modèle $G(n,p)$ peut avoir toutes les arêtes, ou aucune

Quelques résultats

On peut montrer que :

- le degré attendu dans $G(n,p)$ est : $c = (n-1)p$
- le nombre d'arêtes attendu dans $G(n,p)$ est : $\binom{n}{2}$
- le coefficient de clustering attendu pour un nœud (local) est p
- le coefficient de clustering attendu (global) est p
- la longueur moyenne des chemins attendue est $l \sim \frac{ln|V|}{lnc}$
- la probabilité d'avoir m arêtes suit une distribution binomiale :

$$P(|E| = m) = \binom{\binom{n}{2}}{m} p^m (1-p)^{\binom{n}{2}-m}$$

- Quand p s'accroît, une grande partie des nœuds commencent à être interconnectés, on assiste à la formation d'une composante connexe géante

Modélisation avec des graphes aléatoires

- On calcule le degré moyen c du graphe réel
- On calcule une probabilité $p = \frac{c}{n-1}$
- On génère le graphe.

Représentativité

- ✗ ces graphes n'ont pas de distribution en loi de puissance
- ✓ ils sont plutôt bons pour modéliser les longueurs des chemins
- ✗ ils sous-estiment grandement le coefficient de clustering (en général)

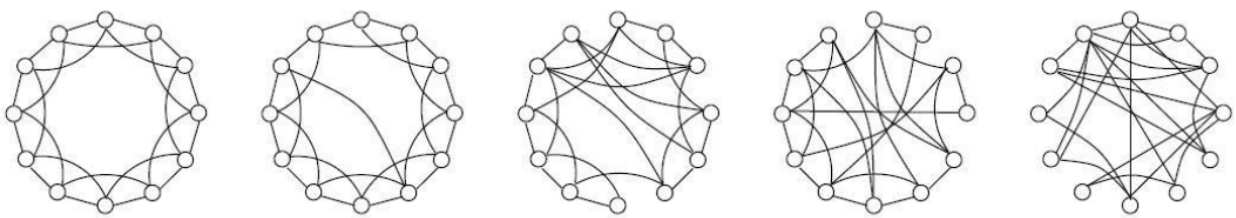
Exemples

	Original Network				Simulated Random Graph	
Network	*Size*	*Average Degree*	*Average Path Length*	C	*Average Path Length*	C
Film Actors	225,226	61	3.65	0.79	2.99	0.00027
Medline Coauthorship	1,520,251	18.1	4.6	0.56	4.91	1.8×10^{-4}
E.Coli	282	7.35	2.9	0.32	3.04	0.026
C.Elegans	282	14	2.65	0.28	2.25	0.05

Modèle petits mondes

- Dans de nombreux réseaux réels, des nœuds ont un nombre réduit, voire fixé, de connexions
- Cela correspond à un graphe régulier (treillis)
- Ce treillis a un coefficient de clustering élevé (mais fixé)
- Ce treillis a une longueur moyenne élevée

- on peut construire un réseau petit monde, paramétré ($0 \leq p \leq 1$) en recâblant les arêtes



Propriétés des réseaux petits mondes générés

- Coefficient de clustering : $C(p) \sim (1-p)^3 C(0)$
- Longueur moyenne des chemins : $l = \frac{n}{2c}$ (élevée)
- Distribution de degré : beaucoup de nœuds avec le même degré (treillis)

Modélisation

- on connaît le degré moyen c et le coefficient de clustering C
- on fixe $C(p) = C$, on estime p à partir de $C(p) \sim (1-p)^3 C(0)$
- on peut modéliser (en ajoutant n)

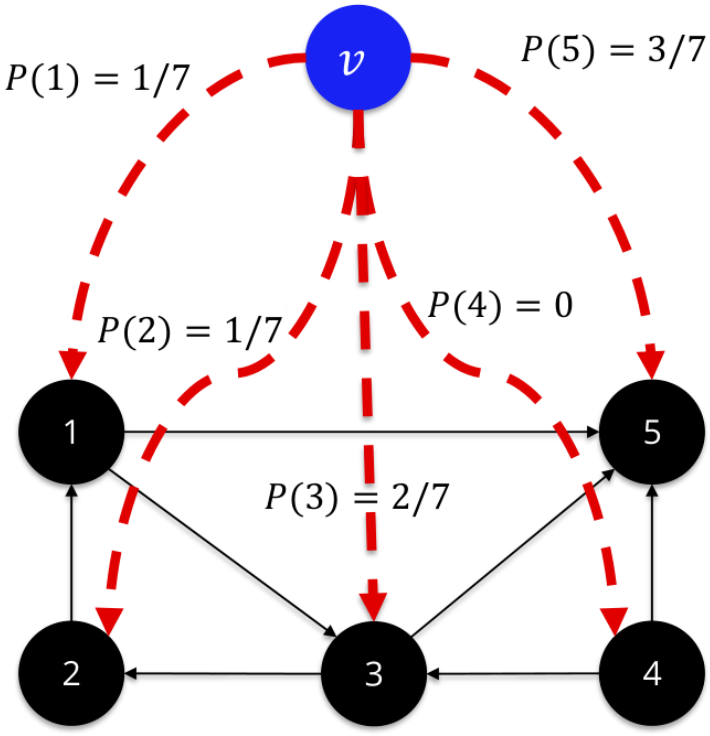
Exemple

	Original Network				Simulated Graph	
Network	*Size*	*Average Degree*	*Average Path Length*	*C*	*Average Path Length*	*C*
Film Actors	225,226	61	3.65	0.79	4.2	0.73
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.1	0.52
E.Coli	282	7.35	2.9	0.32	4.46	0.31
C.Elegans	282	14	2.65	0.28	3.49	0.37

Attachement préférentiel

Modèle d'Albert-Laszlo Barabási et Réka Albert

- Quand un utilisateur rejoint un réseau, la probabilité qu'il se connecte à des nœuds déjà dans le réseau est proportionnelle au degré (entrant) des nœuds
- $P(v_i) = \frac{d_i}{\sum_j d_j}$
- "the rich get richer", "effet Matthieu", "Yule process"



- $P(1) = 1/7$
- $P(2) = 1/7$
- $P(3) = 2/7$
- $P(4) = 0$
- $P(5) = 3/7$

Propriétés, modélisation

- Distribution de degré : $P(d) = \frac{2m^2}{d^3}$
- Longueur des chemins : $l \sim \frac{ln|V|}{ln(ln|V|)}$

Modéliser :

- Comme avec les graphes aléatoires, on peut simuler des graphes réels en fixant le degré espéré m

Exemple

	Original Network				Simulated Graph	
Network	*Size*	*Average Degree*	*Average Path Length*	*C*	*Average Path Length*	*C*
Film Actors	225,226	61	3.65	0.79	4.90	≈ 0.005
Medline Coauthorship	1,520,251	18.1	4.6	0.56	5.36	≈ 0.0002
E.Coli	282	7.35	2.9	0.32	2.37	0.03
C.Elegans	282	14	2.65	0.28	1.99	0.05

Effets inattendus

- Les étapes initiales de l'ascension vers la popularité sont fragiles
- une fois qu'un utilisateur est bien établi dans le réseau, la dynamique de popularité "rich get richer" le pousse plus haut
- mais : obtenir la dynamique "rich get richer" est délicate, pleine d'accidents potentiels et de ratés
- cf : Salganik et al. Science (2006)

Configuration model

- On souhaite avoir un graphe aléatoire avec une distribution de degré donnée

- Étapes, pour une séquence $\{d_1, d_2, ..., d_n\}$:
  1. on génère n nœuds, et le nœud i a d_i demi-arêtes
  2. on attache les demi-arêtes aléatoirement, jusqu'à ce qu'il n'y en ait plus
- il peut y avoir des boucles, des liens multiples
- chaque configuration (mais pas chaque graphe) apparaît avec une égale probabilité

Générer un graphe en configuration model

- on crée une liste, dans laquelle l'id du nœud v_i (degré d_i) est répété d_i fois
- on shuffle la liste
- en partant du premier élément de la liste, on joint les éléments consécutifs deux à deux

- la probabilité qu'un nœud v_i soit connecté à un nœud v_j est :

$$\frac{d_i d_j}{2m-1}$$

  - pour chaque v_i, il y a d_j instances de v_j à côté desquelles il peut être
  - la probabilité d'être proche de l'un d'eux est $\frac{d_j}{(2m-1)}$
  - il y a d_i instances de v_j

Conclusion

Références

Ce cours repose sur les travaux et documents suivants :

- le livre *Social Media Mining* de R. Zafarani, M. A. Abbasi, and H. Liu. Cambridge Univ. Press, 2014. Livre et slides gratuits disponibles sur `http://socialmediamining.info`
- l'équipe ComplexNetworks du LIP6 (Sorbonne Université, `http://www.complexnetworks.fr`), en particulier les cours de Jean-Loup Guillaume (PR, U. de La Rochelle) et de Clémence Magnien (DR CNRS)
- le livre *Mining Massive datasets* (`http://www.mmds.org`), de Jure Leskovec, Anand Rajaraman, Jeff Ullman