innovatives

BIG DATA

Valeur des données

Matrice de Google

Communautés cachées

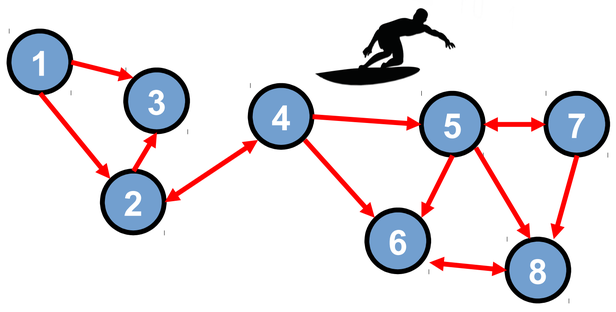
Matrice de Google pour la valorisation des Big Data

José Lages, Dima Shepelyansky, Klaus Frahm, Katia Jaffrès-Runser, Andrei Zinovyev

Un ensemble de mégadonnées peut être vu comme un réseau complexe auquel on peut associer une matrice de Google. De celle-ci, il est possible d'extraire des informations pertinentes concernant ces mégadonnées. Les auteurs de ce poster, participants du projet ApliGoogle lauréat 2016 du défi MASTODONS du CNRS (http://www.quantware.ups-tlse.fr/apligoogle/), ont participé au développement des outils statistiques d'analyse des données issus de la matrice de Google. Parmi ceux-ci, des algorithmes de classement, le CheiRank et le 2DRank (https://en.wikipedia.org/wiki/CheiRank) ont été élaborés, partageant les mêmes fondements théoriques que le PageRank proposé par Brin et Page, cofondateurs de l'entreprise Google. Très récemment, les auteurs de ce poster ont proposé et étudié la matrice de Google réduite associée aux mégadonnées. Cette matrice de Google réduite permet de déceler les communautés cachées dans les mégadonnées. En effet, elle permet de s'intéresser à un sous ensemble d'entités parmi toutes les mégadonnées et permet d'établir des liens cachés (liens à « longues portées ») entre deux entités a priori non directement liées.

De Markov (1906) à Brin & Page (1998)

Processus markovien : un « surfeur » aléatoire sonde la structure du réseau dirigé. A chaque étape, le surfeur choisit aléatoirement un nœud adjacent pour continuer son périple.



Matrice d'adjacence

$$A_{ij} = \begin{cases} 0 & \text{si } j \to i \\ 1 & \text{si } j \nrightarrow i \end{cases}$$

Matrice stochastique

$$S_{ij} = \begin{cases} A_{ij} / \sum_{k=1}^{N} A_{kj} & \text{si } \sum_{k=1}^{N} A_{kj} \neq 0 \\ 1 & \text{sinon} \end{cases}$$

Matrice de Google

$$G_{ij} = \alpha S_{ij} + (1-\alpha)/N \quad \text{avec } 0.5 < \alpha < 1$$

Vecteur PageRank

$$\mathbf{P} = \lim_{n\to\infty} \mathbf{P}^{(n)} = \lim_{n\to\infty} G^n \mathbf{P}^{(0)}$$

$P_i^{(n)}$ est la probabilité que le surfeur aléatoire tombe sur le nœud i après n étapes.

Le plus important nœud du réseau est celui obtenant la plus grande probabilité.
Définition récursive : un nœud est d'autant plus important qu'il est pointé par d'autres nœuds importants.
Le PageRank mesure l'influence d'un nœud.
L'algorithme PageRank est au cœur du moteur de recherche Google

Autre algorithme : l'algorithme CheiRank (Chepelianskii 2010, Zhirov et al, 2010) construit sur le modèle de l'algorithme PageRank mais en considérant le réseau inverse $A^*_{ij} = A_{ji}$. Un nœud est alors d'autant plus important qu'il pointe vers d'autres nœuds importants.

Matrice de Google réduite / liens cachés

Considérons un sous-ensemble de n nœuds, pour lequel on construit la matrice de Google réduite suivante :

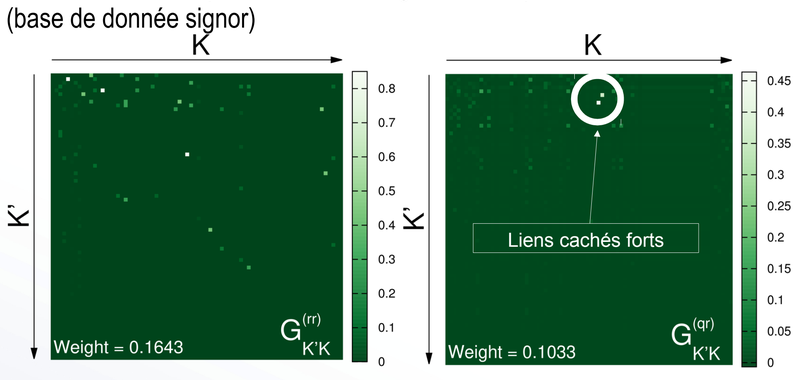
$$G_{(R)} = G_{(rr)} + G_{(pr)} + G_{(qr)}$$

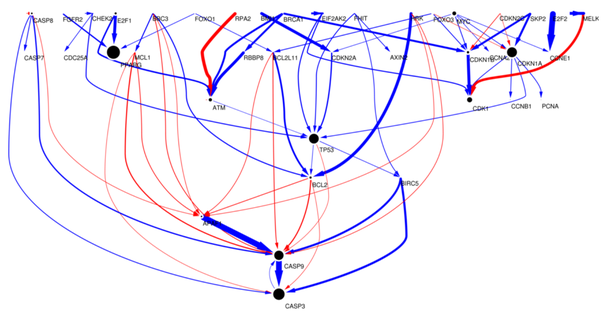
$G_{(R)}$ est telle $G_{(R)} P_{(R)} = P_{(R)}$ où $P_{(R)}$ contient les n éléments du vecteur PageRank de la matrice de Google globale associés aux n nœuds considérés.

$G_{(rr)}$ décrit les liens directs entre les n nœuds / $G_{(pr)}$ encode principalement les informations du PageRank.
$G_{(qr)}$ donne les liens cachés (liens à longue portée) entre les n nœuds considérés.

- Interactions cachées entre protéines en oncologie -

Sous-réseau de 76 protéines / réseau global 2432 protéines
(base de donnée signor)

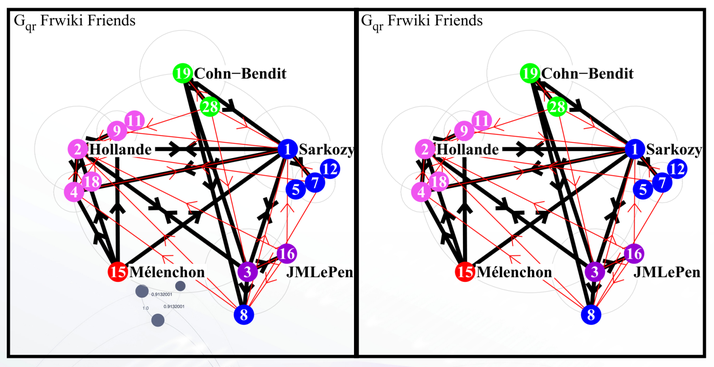




Liens directs en bleu - Liens cachés en rouge

- Liens cachés entre politiciens -

L'étude de la matrice de Google réduite des principaux représentants du corps politique français permet d'établir des influences cachées non évidentes entre politiciens. Pour un politicien donné, il est alors possible de définir des « amis » et des « suiveurs » . Les données ont été obtenues en sondant l'édition française de Wikipédia (2013).



Exemples de résultats obtenus par l'analyse de la matrice de Google

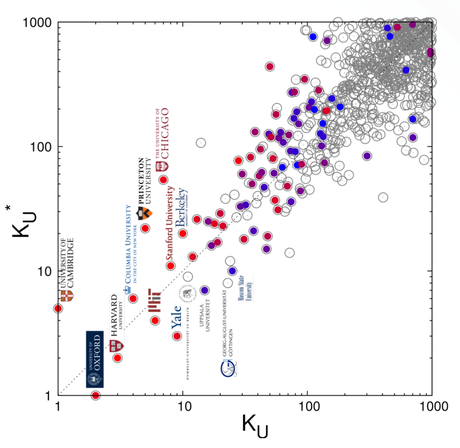
[pour une revue exhaustive voir L. Ermann, K.M. Frahm, D.L. Shepelyansky, Rev. Mod. Phys. 87, 1261 (2015)]

- Classement mondial des universités -

24 éditions Wikipédia (2013) sondées
= 17 millions de d'articles = 59 % de la population mondiale = 68 % des articles Wikipédia
1024 universités ont été classées via PageRank, CheiRank et 2DRank

Wikipedia Ranking of World Universities WRWU		Academic Ranking of World Universities ARWU ("Shanghai ranking" 2013)
1st University of Cambridge	**90% de recouvrement** entre les top 10s WRWU et ARWU	1st Harvard University (-2)
2nd University of Oxford		2nd Stanford University (-6)
3rd Harvard University		3rd University of California, Berkeley (-7)
4th Columbia University		4th MIT (-2)
5th Princeton University		5th University of Cambridge (+4)
6th MIT	**60% de recouvrement** entre les top 100s WRWU et ARWU	6th California Institute of Technology (-22)
7th University of Chicago		7th Princeton University (+2)
8th Stanford University		8th Columbia University (+4)
9th Yale University		9th University of Chicago (+2)
10th University of California, Berkeley		10th University of Oxford (+8)

A l'instar du classement ARWU (dit de « Shanghai »), le classement WRWU mesure l'excellence académique, mais aussi l'influence historique et régionale des universités.

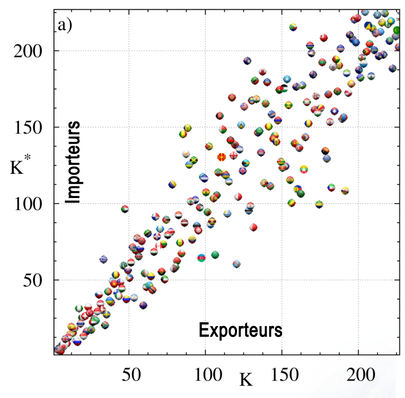


- Matrice de Google du commerce international -

L'analyse de la matrice de Google du commerce international permet de traiter tous les pays du monde, riches ou pauvres, sur un même pied d'égalité. Pour un même pays, les quantités importées ou exportées sont rapportées aux quantités totales importées ou exportées.

Le PageRank (CheiRank) va naturellement caractériser la propension d'un pays à exporter (importer).

Les données utilisées sont celle de l'ONU (UN COMTRADE).



Mégadonnées vues comme des réseaux dirigés complexes

Données	Nœuds	Liens entre nœuds
WWW	Pages web	Hyperliens
Wikipédia	Articles	Citations intra-wiki
Twitter	Membres	Follower
Commerce International	Produits économiques	Balance économique entre pays par produits
Omiques	Protéines	Inhibition/activation
Linux	Commandes du noyau	Succession des commandes
ADN	Séquences	Succession des séquences
Actitvité cérébrale	Neurones	Connexions synaptiques
Jeu de Go	Motifs joués	Succession des motifs

Exemples de mégadonnées étudiées

Références : Google matrix analysis of directed networks, L. Ermann, K.M. Frahm, D.L. Shepelyansky, Rev. Mod. Phys. 87, 1261 (2015) / Wikipedia Ranking of World Universities, J. Lages, A. Patt, D. L. Shepelyansky, Eur. Phys. J. B (2016) 89: 69 / Google matrix of the world trade network, L. Ermann, D.L. Shepelyansky, Acta Physica Polonica A, vol. 120 (6A), A-158 (2011) / Google matrix analysis of cancer protein networks, J. Lages, D.L. Shepelyansky, A. Zinovyev, en préparation / Wikipedia mining of hidden links between political leaders, Kl.M. Frahm, K. Jaffrès-Runser, D.L. Shepelyansky, arXiv:1609.01948

Contact : jose.lages@utinam.cnrs.fr Institut UTINAM / UMR CNRS 6213 / Université de Franche-Comté

www.cnrs.fr

ApliGoogle
Applications de la matrice Google pour les réseaux directionnels et Big Data

UFC UNIVERSITÉ DE FRANCHE-COMTÉ

UNIVERSITÉ TOULOUSE III PAUL SABATIER Université de Toulouse

institutCurie

Inserm

Institut UTINAM

LPT LABORATOIRE DE PHYSIQUE THÉORIQUE TOULOUSE

IRIT CNRS - INPT - UPS - UT1 - UT2J