Révolution numérique et changements sociétaux

Comment les algorithmes impactent notre manière de nous informer ?

Fabien Tarissan

CNRS – ENS Paris Saclay – ISP

Université Populaire d'Antony

Internet et le web

Internet

1966 projet Arpanet

29 oct. 1969 premier message

↦ *réseau de 4 sommets*

1971 projet Cyclade

1973 protocole TCP/IP (Vint Cerf & Robert Kahn)

↦ *réseaux de ∼ 50 de sommets*

1er jan. 1983 Arpanet adopte TCP/IP

↦ *réseaux de ∼ 1000 de sommets*

1987 Internet dépasse les 20 000 routeurs.

Le web

mars 1989 projet *World Wide Web* du CERN (Tim Berners-Lee & Robert Cailliau)

20 déc. 1990 1ère page web

30 avril 1993 CERN renonce aux droits

↦ *600 pages webs*

1994 *Netscape* et *Yahoo !*

1998 *Google*

2000 – ... *Facebook*, *Youtube*, *Twitter*, ...

2007 – ... *Deezer*, *Spotify*, *Netflix*, ...

En quête de visibilité

> *Tous les deux jours, nous créons autant d'information que l'humanité tout entière entre l'aube de son histoire et l'année 2003.*
>
> Eric Schmidt (PDG Google) en 2010

Le web en chiffres (2022) `http://www.internetlivestats.com/`

- 5,5 milliards d'utilisateur
- 1,9 milliards de sites web
- 500 millions de tweets et 5 milliards de vidéos par jours
- 5 milliards de requêtes et 200 milliards de mails envoyés par jours

$\Longrightarrow$ Nécessité d'algorithmes de classement

Quel impact sur l'information rendue visible ?

1. Les algorithmes de classement
2. L'impact des algorithmes sur notre manière de nous informer

Moteurs de recherche

Moteurs de recherche

Tâche d'un moteur de recherche :

- Sur réception d'une liste de mots-clefs (ex : "science" et "réseaux")
- Renvoyer une liste de pages web pertinentes *vis-à-vis* de ces mots-clefs.
  1. `https://fr.wikipedia.org/wiki/Science_des_reseaux`
  2. `https://www.pourlascience.fr/sd/informatique/le-monde-a-ses-reseaux-1082.php`
  3. ...

Moteurs de recherche

Tâche d'un moteur de recherche :

- Sur réception d'une liste de mots-clefs (ex : "science" et "réseaux")
- Renvoyer une liste de pages web pertinentes *vis-à-vis* de ces mots-clefs.
  1. `https://fr.wikipedia.org/wiki/Science_des_reseaux`
  2. `https://www.pourlascience.fr/sd/informatique/le-monde-a-ses-reseaux-1082.php`
  3. ...

Plus dur qu'il n'y paraît :

Collecte : Quelle est l'information disponible ?
Connaître le contenu et l'emplacement de l'ensemble des pages web

$\mapsto$ fait en continu par des programmes routiniers (*crawlers* ou *spider*)
(voir `https://fr.wikipedia.org/robots.txt`, `https://www.lemonde.fr/robots.txt`)

Indexation : Quelles pages correspondent aux requêtes ?
Génération d'une indexation complète associant à chaque mot-clef une liste de page web. $\mapsto$ mis à jour en continu

Classement : Comment ordonner les pages ?
Choix d'un critère de classement : comment quantifier la pertinence ?
Algorithme pour systématiser le calcul

$\mapsto$ fait en continu

L'importance d'une page web

Question (1) : comment quantifier l'importance d'une page web ?

L'importance d'une page web

Question (1) : comment quantifier l'importance d'une page web ?

- Avant Google : médiamétrie (traffic, nb de visites, de clics, ...) $\Longrightarrow$ popularité

L'importance d'une page web

Question (1) : comment quantifier l'importance d'une page web ?

- Avant Google : médiamétrie (traffic, nb de visites, de clics, ...) $\Longrightarrow$ popularité
- Idée Brin & Page (1998) : les références hypertextes constituent des gages d'autorité

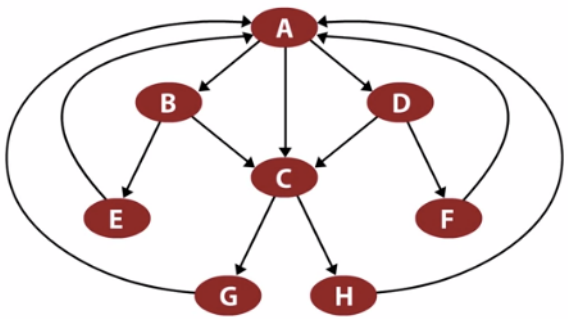
L'importance d'une page web

Question (1) : comment quantifier l'importance d'une page web ?

- Avant Google : médiamétrie (traffic, nb de visites, de clics, ...) $\Longrightarrow$ popularité
- Idée Brin & Page (1998) : les références hypertextes constituent des gages d'autorité

Un des résultats de la phase de collecte : le réseau du web, ie. un graphe :

- Nœuds = pages web
- Liens = liens hypertextes



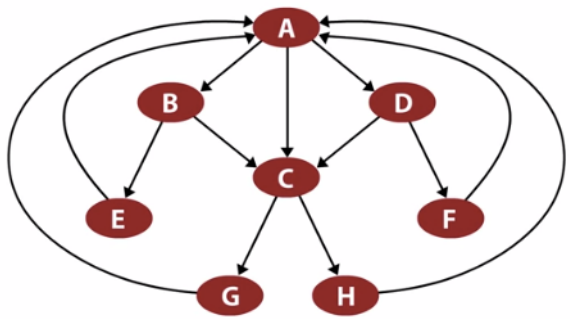
L'importance d'une page web

Question (1) : comment quantifier l'importance d'une page web ?

- Avant Google : médiamétrie (traffic, nb de visites, de clics, ...) $\Longrightarrow$ popularité
- Idée Brin & Page (1998) : les références hypertextes constituent des gages d'autorité

Un des résultats de la phase de collecte : le réseau du web, ie. un graphe :

- Nœuds = pages web
- Liens = liens hypertextes



Question (2) : comment quantifier l'importance d'un nœud dans un réseau de citations ?

PageRank et marches aléatoires

Pagerank est basée sur la notion de marche aléatoire :

1. un marcheur part d'un nœud choisi aléatoirement
2. il choisit aléatoirement l'un des liens sortants du nœud
3. retourne en 2

PageRank et marches aléatoires

Pagerank est basée sur la notion de marche aléatoire :

1. un marcheur part d'un nœud choisi aléatoirement
2. il choisit aléatoirement l'un des liens sortants du nœud
3. retourne en 2

Importance (score) d'un nœud v = probabilité d'observer le marcheur aléatoire sur le nœud v après une infinité d'étapes.

PageRank et marches aléatoires

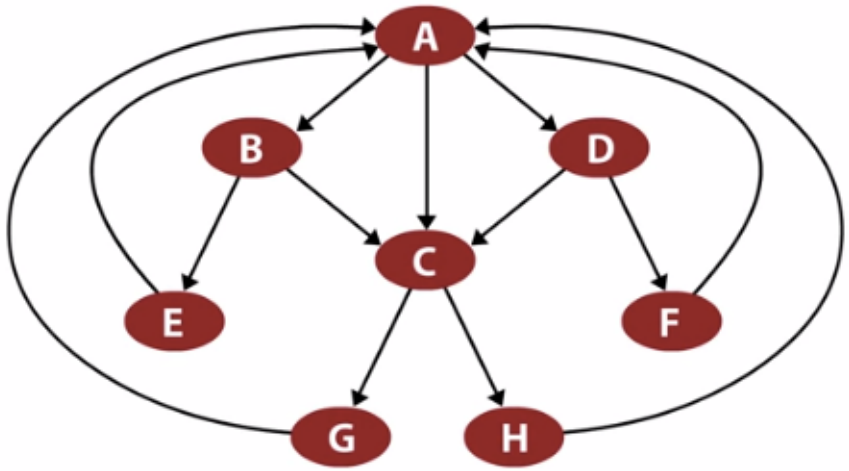
Pagerank est basée sur la notion de marche aléatoire :

1. un marcheur part d'un nœud choisi aléatoirement
2. il choisit aléatoirement l'un des liens sortants du nœud
3. retourne en 2

Importance (score) d'un nœud v = probabilité d'observer le marcheur aléatoire sur le nœud v après une infinité d'étapes.

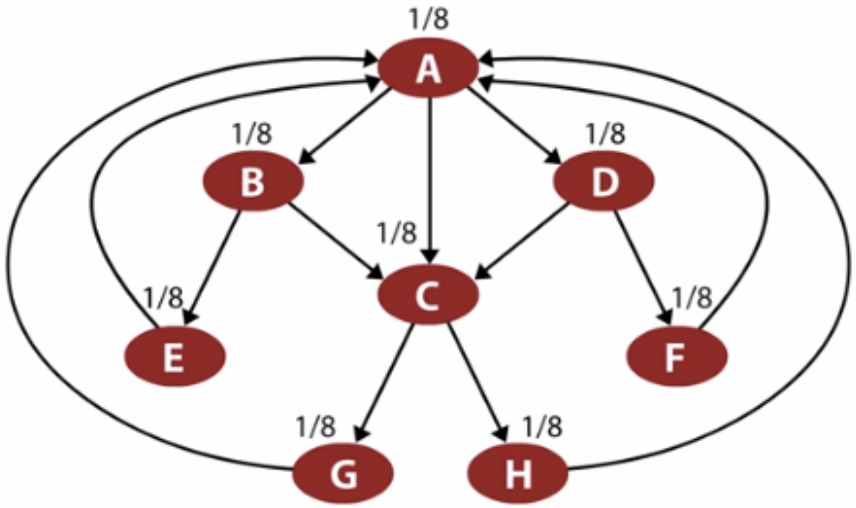
Question (3) : comment estimer cette probabilité ?

Algorithme du PageRank : intuition



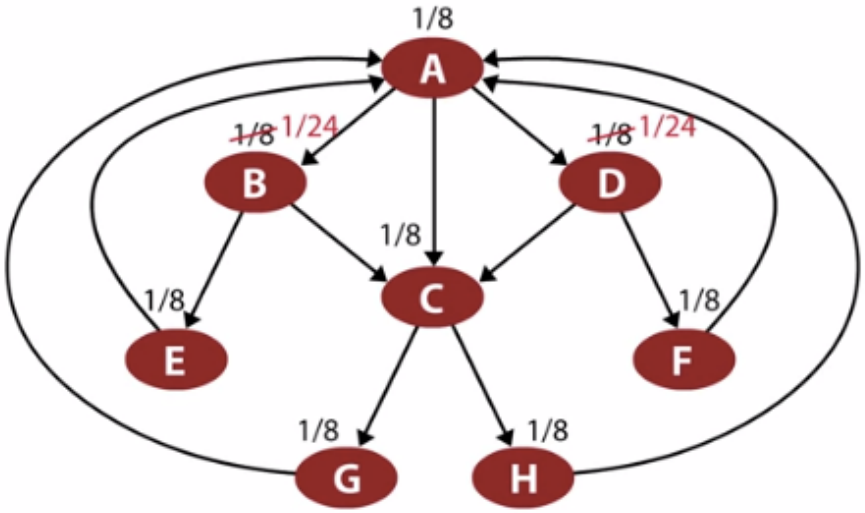
The PageRank Citation Ranking : Bringing Order to the Web, Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, Technical Report. Stanford InfoLab, 1998.

Algorithme du PageRank : intuition



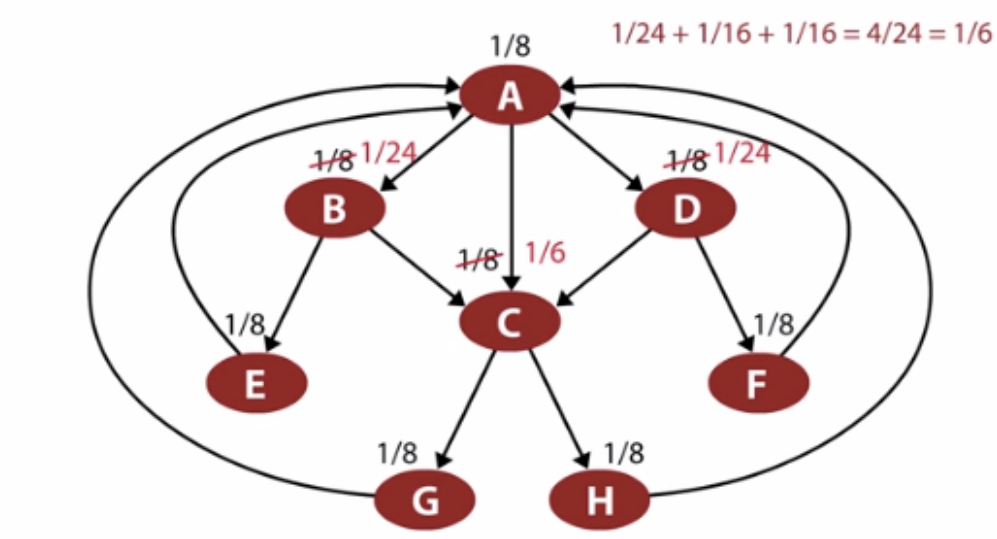
On donne un poids équivalent (normalisé) à l'ensemble des sommets du graphe : 1/N

Algorithme du PageRank : intuition



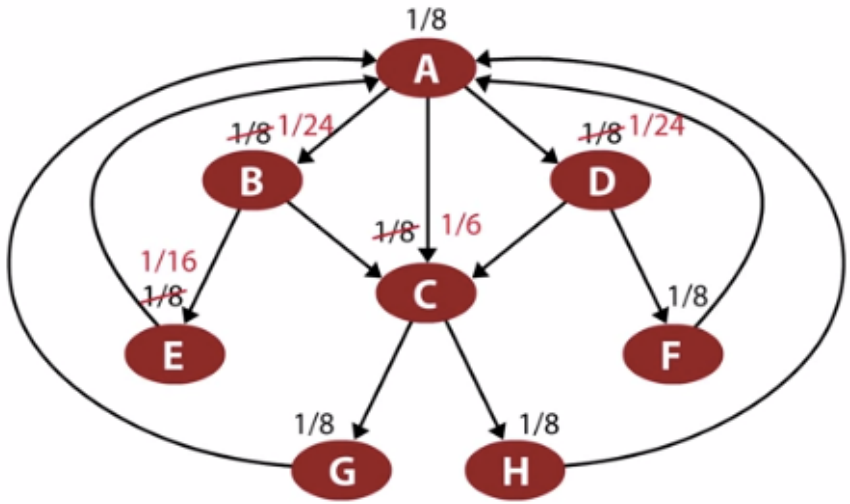
MAJ : On répartit simultanément le PR équitablement vers les nœuds sortants

Algorithme du PageRank : intuition



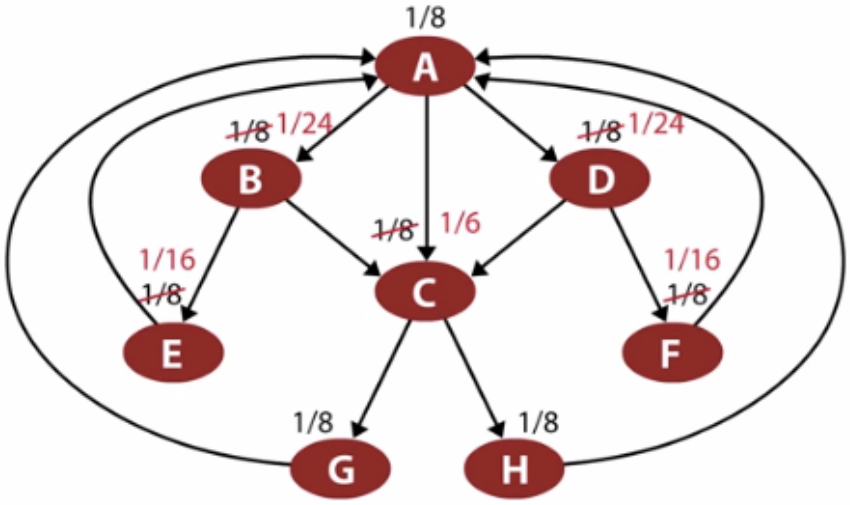
MAJ : On répartit simultanément le PR équitablement vers les nœuds sortants

Algorithme du PageRank : intuition



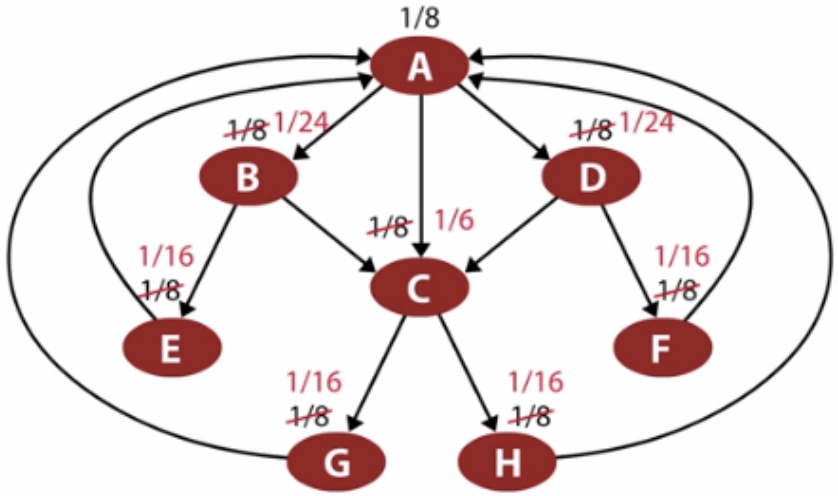
MAJ : On répartit simultanément le PR équitablement vers les nœuds sortants

Algorithme du PageRank : intuition



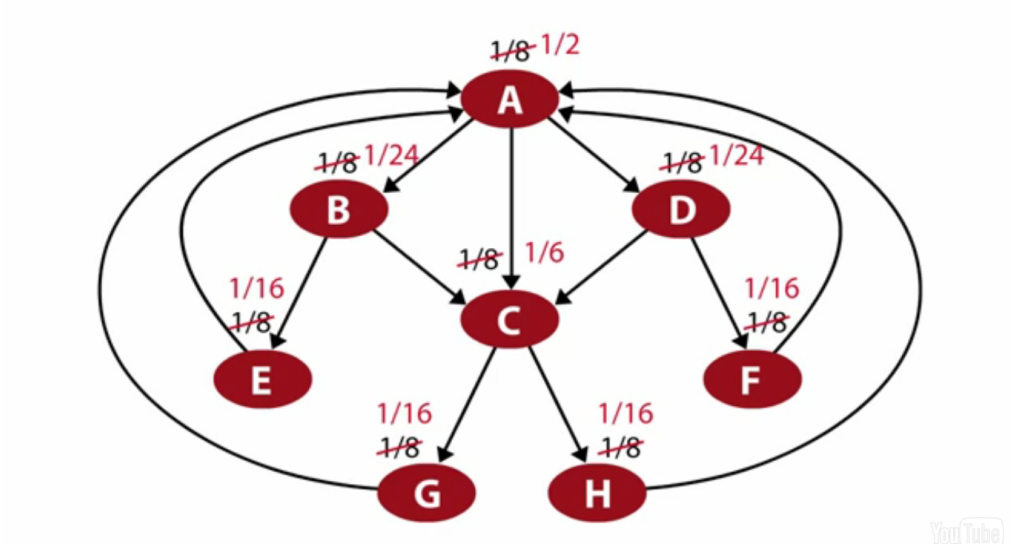
MAJ : On répartit simultanément le PR équitablement vers les nœuds sortants

Algorithme du PageRank : intuition



MAJ : On répartit simultanément le PR équitablement vers les nœuds sortants

Algorithme du PageRank : intuition

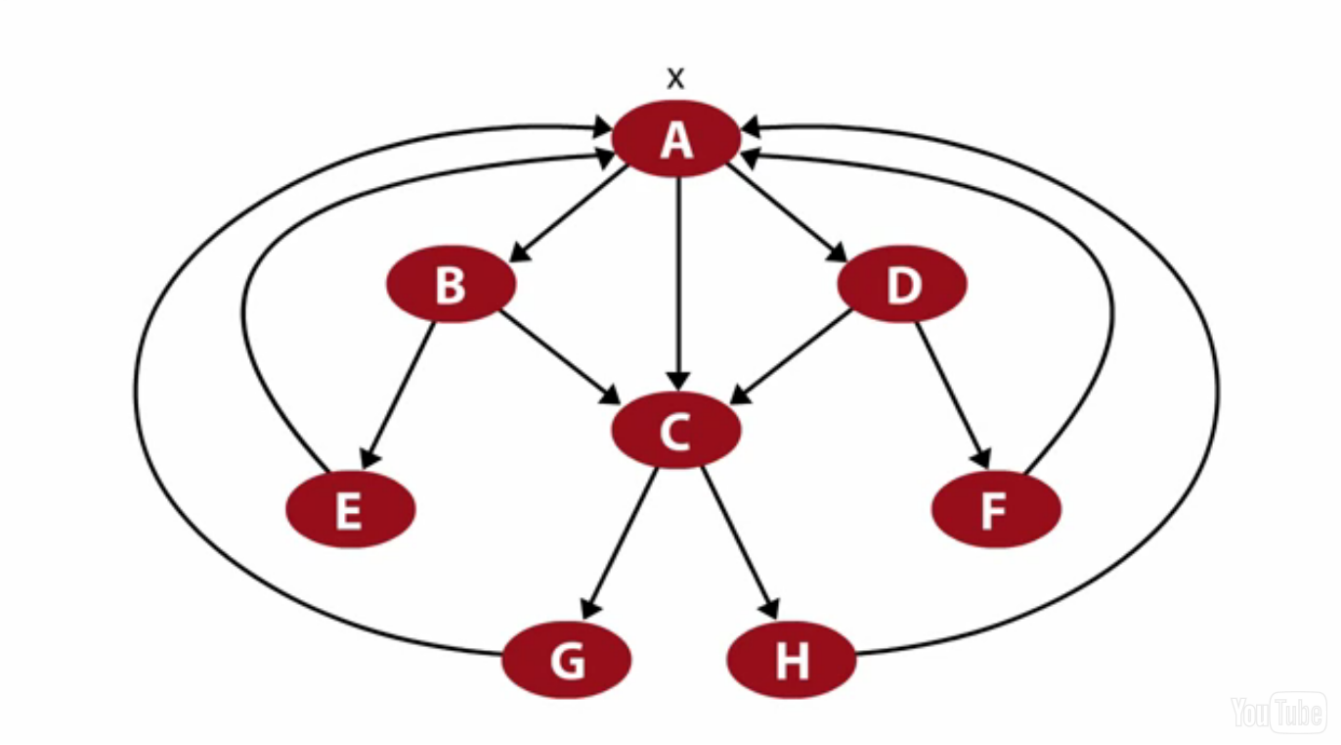


MAJ : On répartit simultanément le PR équitablement vers les nœuds sortants

Algorithme du PageRank : principe

Quel est l'état final ?

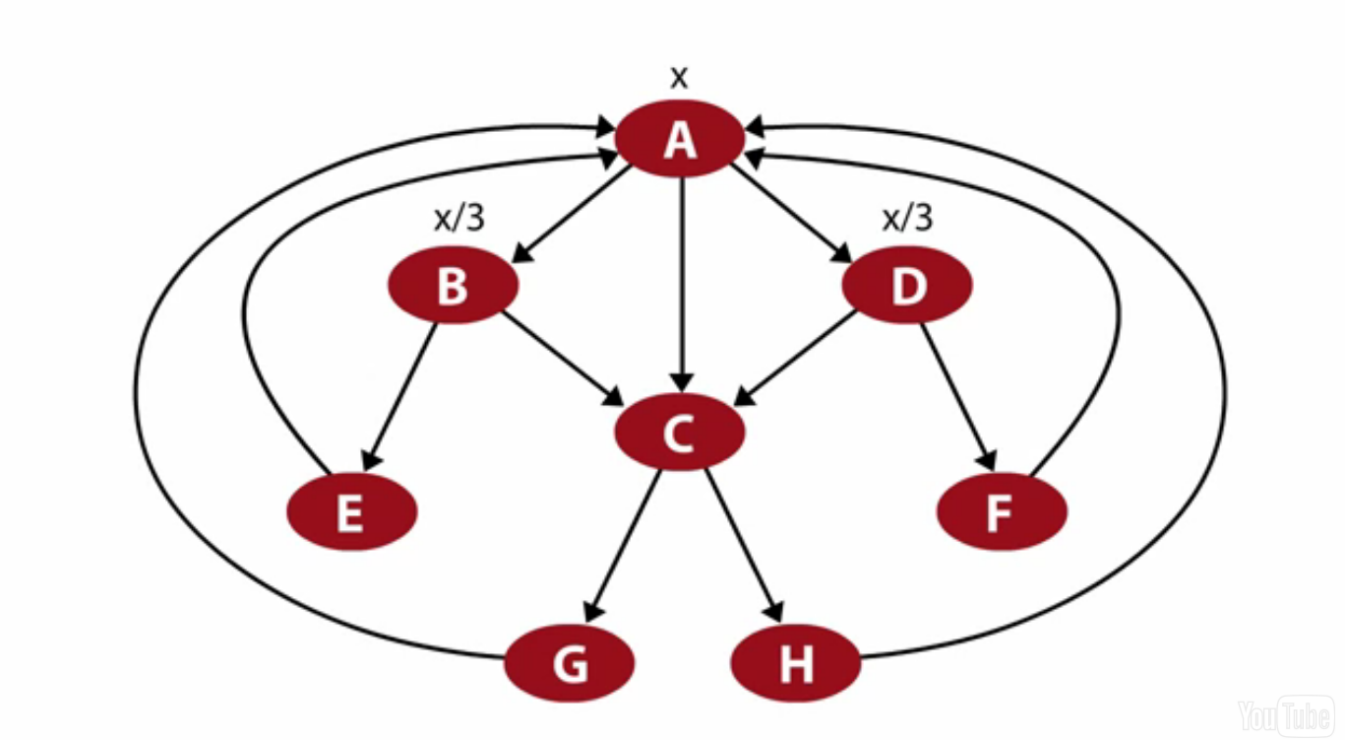
[admis] un seul état d'équilibre



Algorithme du PageRank : principe

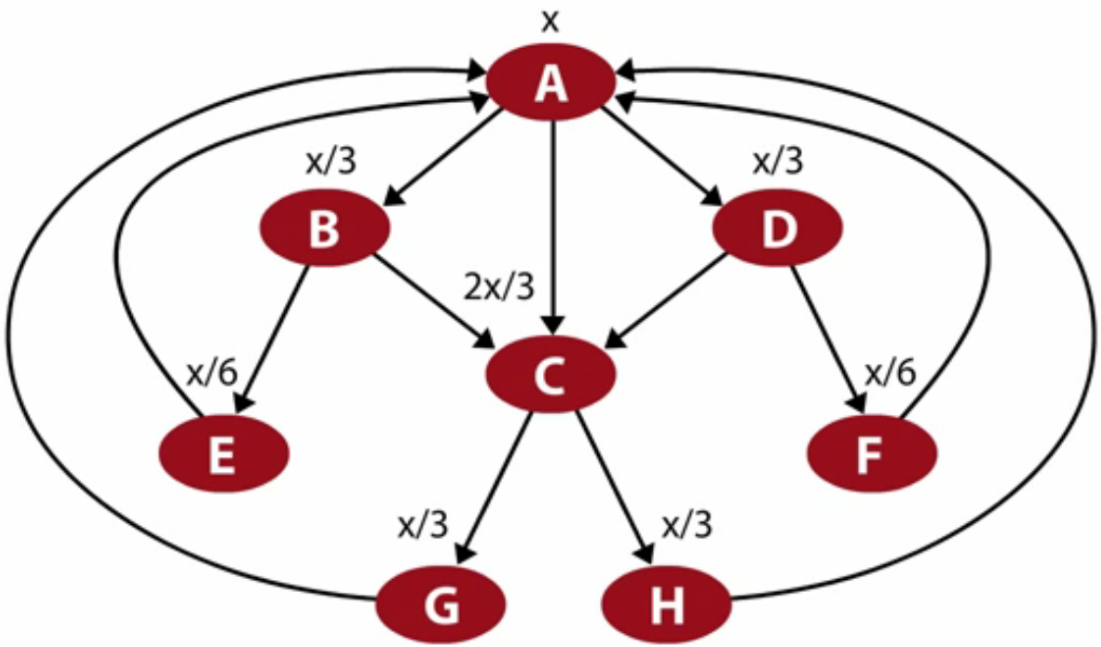
Quel est l'état final ?

[admis] un seul état d'équilibre



Algorithme du PageRank : principe

Quel est l'état final ?
[admis] un seul état d'équilibre



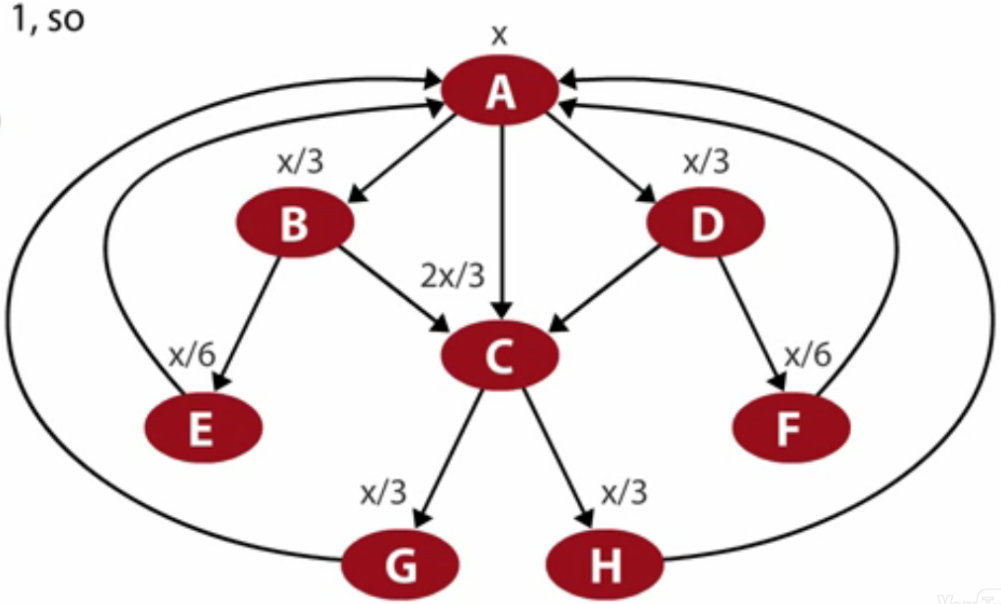
Algorithme du PageRank : principe

Quel est l'état final ?
[admis] un seul état d'équilibre

$x + x/3 + x/3 + 2x/3 + x/6 + x/6 + x/3 + x/3 = 10x/3$

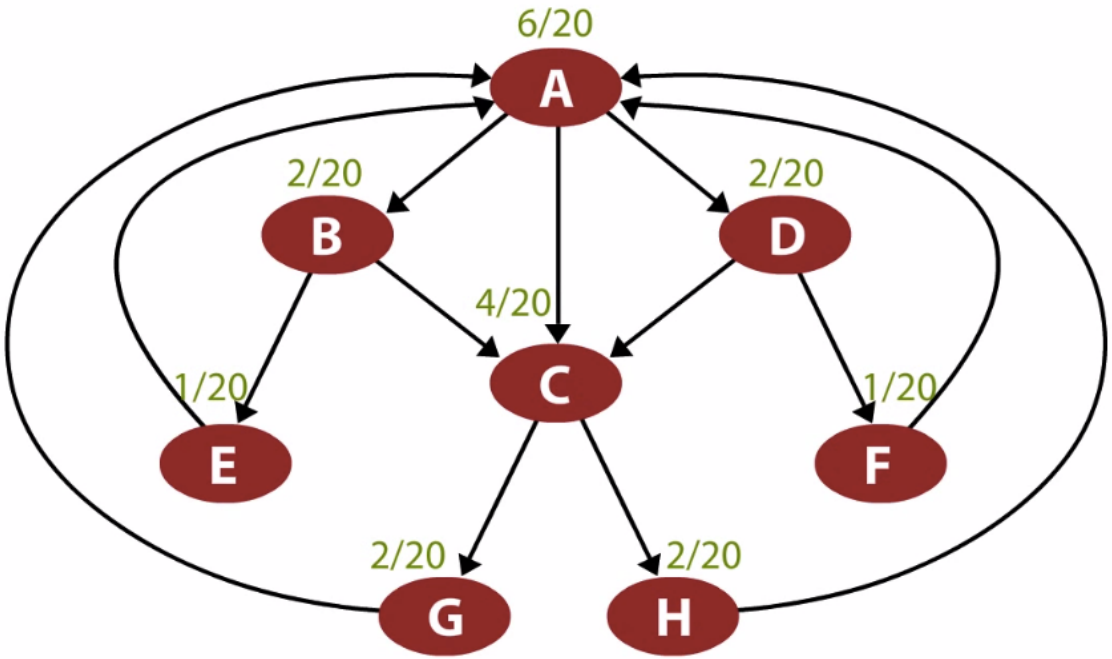
This must equal 1, so

- $10x/3 = 1$
- $x = 3/10 = 6/20$



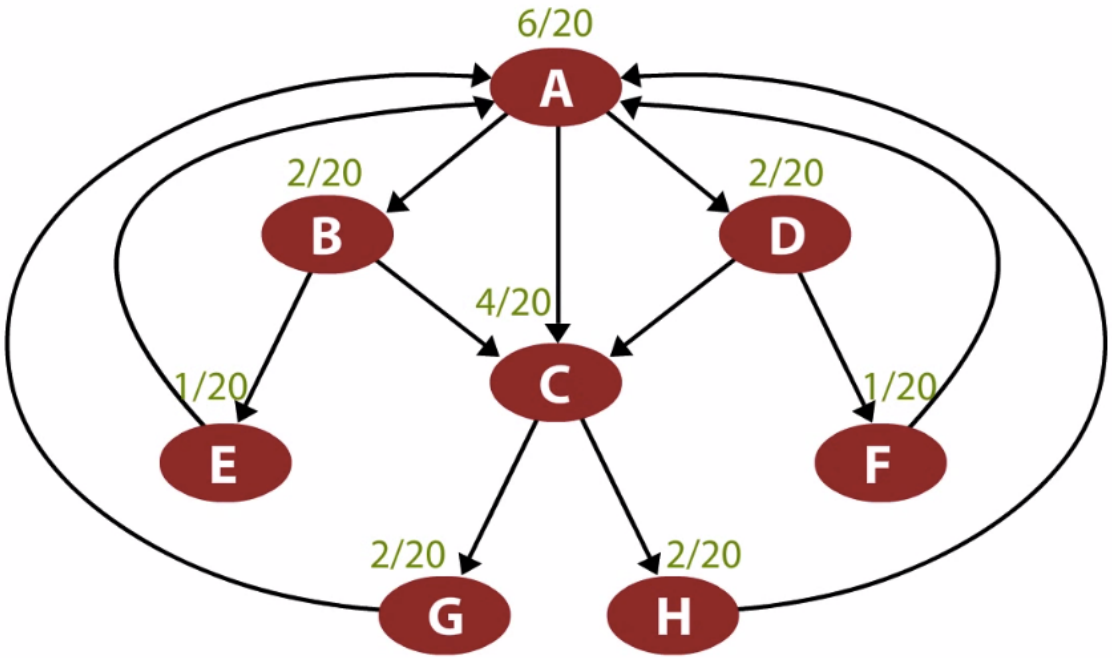
Algorithme du PageRank : principe

Quel est l'état final ?
[admis] un seul état d'équilibre



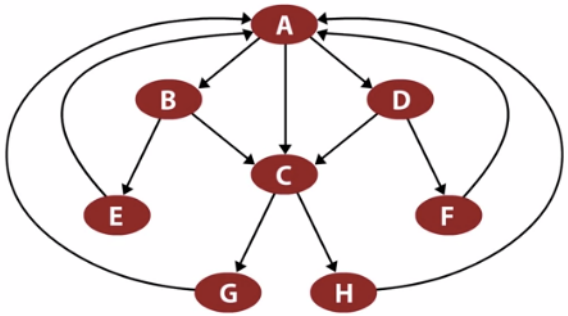
Algorithme du PageRank : principe

Quel est l'état final ?
[admis] un seul état d'équilibre



Question (4) : comment déterminer les valeurs stationnaires ?

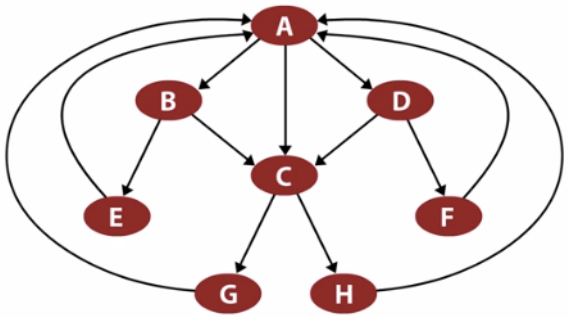
Algorithme du PageRank : matrices



On cherche à résoudre le système d'équations suivant :

- $PR_A = PR_F + PR_H + PR_G + PR_E$
- $PR_B = \frac{PR_A}{3}$
- $PR_C = \frac{PR_A}{3} + \frac{PR_B}{2} + \frac{PR_D}{2}$
- ...
- $PR_H = \frac{PR_C}{2}$

Algorithme du PageRank : matrices

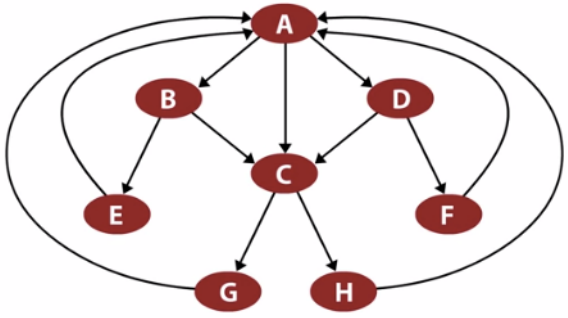


On cherche à résoudre le système d'équations suivant :

- $PR_A = PR_F + PR_H + PR_G + PR_E$
- $PR_B = \frac{PR_A}{3}$
- $PR_C = \frac{PR_A}{3} + \frac{PR_B}{2} + \frac{PR_D}{2}$
- ...
- $PR_H = \frac{PR_C}{2}$

$\Longrightarrow$ C'est à dire à déterminer le vecteur $PR = [PR_A, PR_B, ..., PR_H]$

Algorithme du PageRank : matrices

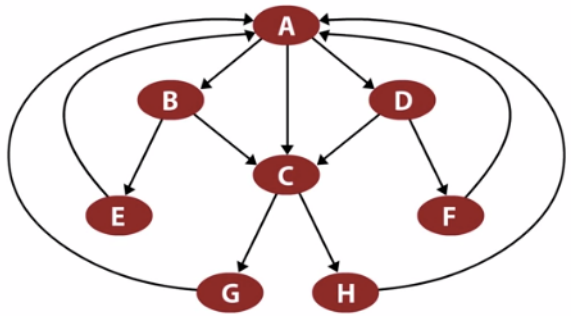


Matrice de transition T du graphe G :

- matrice carré $n \times n$
- chaque lien dirigé (u, v) de G donne la valeur $T[u][v] = \frac{1}{deg(u)}$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 1/3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Algorithme du PageRank : matrices

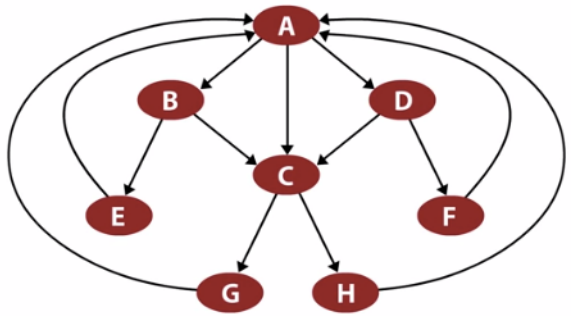


Si PR est le vecteur contenant le score du PageRank (inconnue), on retrouve le système d'équation :

$$T \times PR = PR$$

PR est un vecteur propre de la matrice de transition du graphe G.

Algorithme du PageRank : matrices



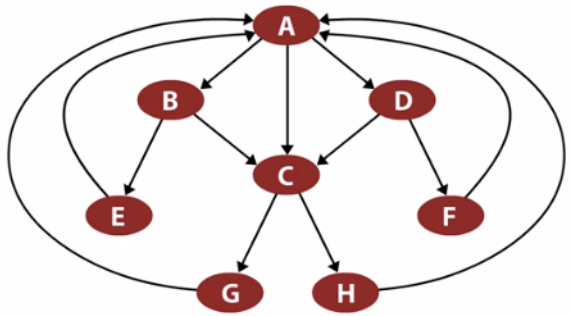
Si PR est le vecteur contenant le score du PageRank (inconnue), on retrouve le système d'équation :

$$T \times PR = PR$$

PR est un vecteur propre de la matrice de transition du graphe G.

Question (5) : comment calculer un vecteur propre d'une matrice ?

Algorithme du PageRank : matrices



Si PR est le vecteur contenant le score du PageRank (inconnue), on retrouve le système d'équation :

$$T \times PR = PR$$

PR est un vecteur propre de la matrice de transition du graphe G.

Question (5) : comment calculer un vecteur propre d'une matrice ? ... j'arrête !

Réseaux sociaux

Réseaux sociaux

Activités des utilisateurs variées :

- recherche
- lecture
- dépôts (*posts*, *tweets*, ...)
- marques d'intérêt (commentaires, *likes*, ...)
- relations ("amis" FB, *followers*, ...)
- propagation (*retweets*, ...)
- ...

Réseaux sociaux

Activités des utilisateurs variées :

- recherche
- lecture
- dépôts (*posts*, *tweets*, ...)
- marques d'intérêt (commentaires, *likes*, ...)
- relations ("amis" FB, *followers*, ...)
- propagation (*retweets*, ...)
- ...

Tâche des réseaux sociaux :

Identifier les informations pertinentes

- adaptées à chaque utilisateurs
- sans être guidé par des mot-clefs

Exemple de Facebook : le *Newsfeed*

Pari de Zuckerberg (2004) : c'est la régularité de l'activité des utilisateurs qui va guider l'algorithme pour identifier leurs centres d'intérêt.

EdgeRank

Comme pour un moteur de recherche, il y a plusieurs phases

Collecte : récolter les informations susceptibles d'intéresser un utilisateurs

$\mapsto$ relations entre utilisateurs ($\sim$ facile)

Classement : ordonner les posts

$\mapsto$ l'algorithme du EdgeRank

EdgeRank

Comme pour un moteur de recherche, il y a plusieurs phases

Collecte : récolter les informations susceptibles d'intéresser un utilisateurs

$\mapsto$ relations entre utilisateurs ($\sim$ facile)

Classement : ordonner les posts

$\mapsto$ l'algorithme du EdgeRank

Edgerank

Chaque relation entre un utilisateur B et un post p d'un "ami" A de B est évalué :

$$S(A, B, p) = Aff(B, A) \times W(p) \times D(p)$$

Affinité : $Aff(B, A)$ mesure l'affinité déclarée de B vers A

Poids : $W(p)$ mesure le poids du post p

Fraicheur : $D(p)$ mesure l'ancienneté du post p

EdgeRank

Comme pour un moteur de recherche, il y a plusieurs phases

Collecte : récolter les informations susceptibles d'intéresser un utilisateurs

$\mapsto$ relations entre utilisateurs ($\sim$ facile)

Classement : ordonner les posts

$\mapsto$ l'algorithme du EdgeRank

Edgerank

Chaque relation entre un utilisateur B et un post p d'un "ami" A de B est évalué :

$$S(A, B, p) = Aff(B, A) \times W(p) \times D(p)$$

Affinité : $Aff(B, A)$ mesure l'affinité déclarée de B vers A

Poids : $W(p)$ mesure le poids du post p

Fraicheur : $D(p)$ mesure l'ancienneté du post p

En fait, EdgeRank est une somme :

$$ER(p) = \sum_B S(A, B, p)$$

Quelle évolution ?

Dominique Cardon, À quoi rêvent les algorithmes. Nos vies à l'heure des big data, Paris, Seuil, La République des idées, 2015, 105 p.

Évolution des algorithmes classements de l'information :

Popularité (À côté du web) : nombre de réferences.
Clics, views, ..

Autorité (Au dessus du web) : authorité des réferences
Links $\Longrightarrow$ PageRank

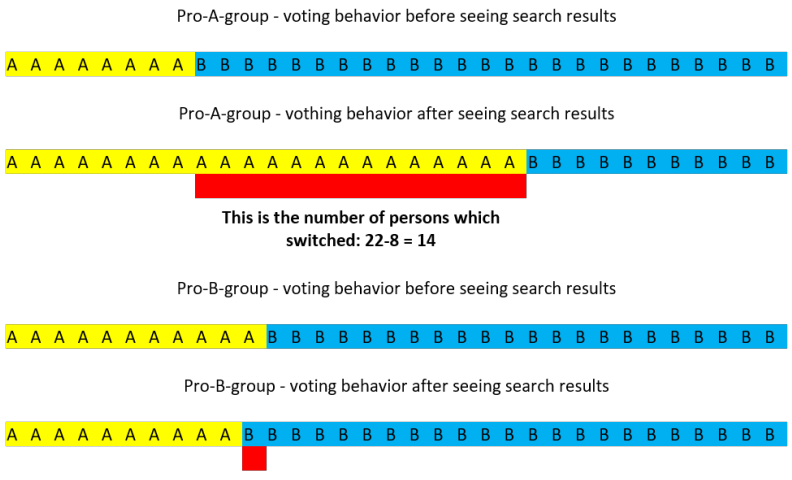
Notoriété (À l'intérieur du web) : affinité (et fraicheur !) avec l'information issue de son voisinage.
Tweet, likes, ... $\Longrightarrow$ EdgeRank

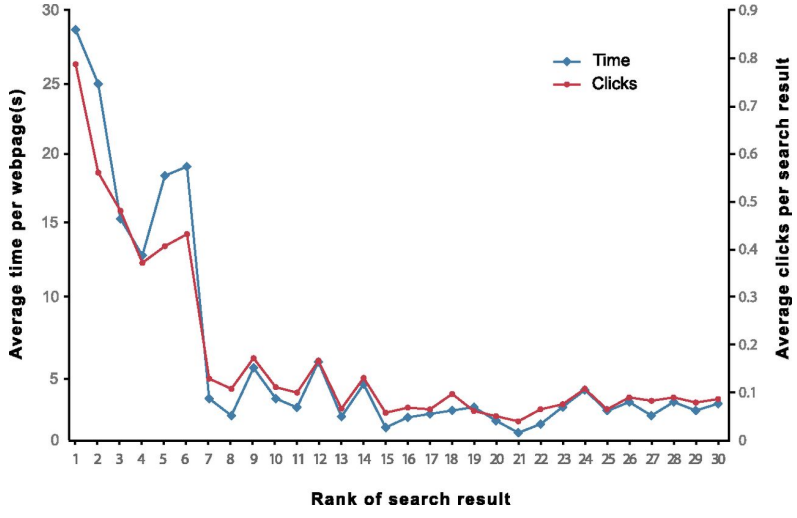
Prédiction (Sous le web) : information inférée par le comportement d'un utilisateur.
Traces $\Longrightarrow$ Youtube / Spotify / Deezer / Netflix / ...

Quel impact ont ces algorithmes

Search Engine Manipulation effect

The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Robert Epstein and Ronald E. Robertson. PNAS 112 (33), 2015 (doi :10.1073/pnas.1419828112).





Résultats

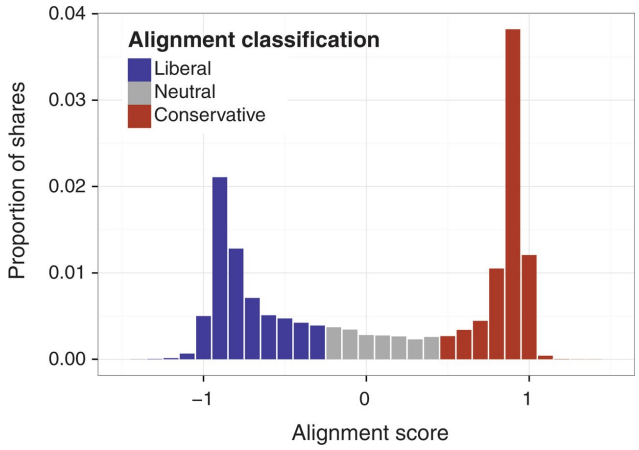
- Classements biaisés ont un impact sur les votants (indécis) : $\geq 25\%$
- Les moteurs de recherche biaisés sont indétectés

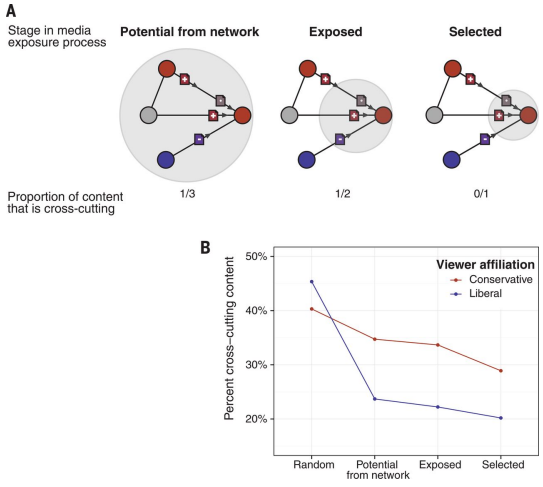
À mettre en perspective avec :

- 95% de la navigation web se fait sur 0.3% des pages existantes ...
- **The Fake News Machine. How Propagandists Abuse the Internet and Manipulate the Public**, Gu, Kropotov and Yarochkin, Trendlabs research paper, Trend Micro, 2015.

Chambres d'écho & bulles filtrantes

Exposure to ideologically diverse news and opinion on Facebook. Eytan Bakshy, Solomon Messing, Lada A. Adamic. Science, 348 (6239), 2015 (doi :10.1126/science.aaa1160).





Résultats (N = 10.1 M)

- Les chambres d'échos sont dues principalement aux relations sociales
- Petit effet du filtrage algorithmique

Quels remparts ?

Science / outils / bonnes pratiques

Prenons l'exemple de l'anonymisation ou des *cookies* :

- Qwant, DuckDuckGo, Brave, VPN, Protonmail, Signal, ...

Insuffisant : Yves-Alexandre de Montjoye, César Hidalgo, Michel Verleysen et Vincent D. Blondel, *Unique in the crowd : The privacy bounds of human mobility*, Scientific Reports, vol. 3, 2013.

Légiférer / réguler / encadrer

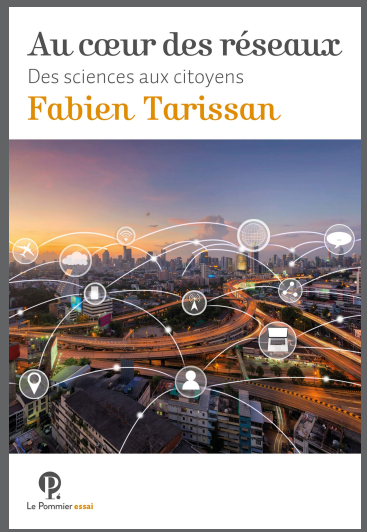
RGPD (2016), Loi pour une République numérique (2016), Loi relative à la lutte contre la manipulation d'information (2018), le DSA/DMA (Europe), la CNIL, ...

Éducation / formation / médiation

Réforme du bac (2019)

- Sciences Numériques et Technologie (SNT) : toutes les 2ndes
- Numérique et Sciences Informatiques (ISN) : spécialité de 1ère et Terminale

Enjeux : **nouveau corps enseignant ?**



Questions ?

http://tarissan.complexnetworks.fr/