

Graphes, Réseaux sociaux

L3 Info

UE Impact environnemental et sociétal du numérique

Julien Lefèvre, Pierre-Alain Reynier

Aix-Marseille Université (AMU)

julien.lefevre@univ-amu.fr, pierre-alain.reynier@univ-amu.fr

Janvier 2022

Contenu

- 1 Introduction
- 2 Analyse de réseaux sociaux
- 3 Modèles de réseaux sociaux
- 4 PageRank

1 - Introduction

Motivations : quelques faits

- Nous vivons une période de **changement climatique** et de **déclin de la biodiversité**.
Les phénomènes précédents sont d'**origine anthropique**.
- Le numérique, qui inclut l'informatique, a une empreinte environnementale, même s'il est difficile de la calculer globalement.
- Le numérique a engendré des bouleversements importants sur nos modes de vie, avec des opportunités (partage de la connaissance...) et des risques (privacy, bulles de filtre, addictions...).
- Les GAFAM ont une place prépondérante dans l'économie mondiale (2.5 fois la valorisation du CAC40).

Différentes phases

Que retenir de l'histoire du numérique ?

<1945 Fondements de l'informatique du côté des mathématiques et de la logique (Von Neumann, Turing)

45-95 Des centres de calcul aux ordinateurs personnels

95-10 **La révolution Internet**

2010- Le "tout connecté" (et la préoccupation environnementale)

2012- **Les neurones contre-attaquent** (l'IA 2.0 : les réseaux profonds)

Digital Revolution, wikipedia

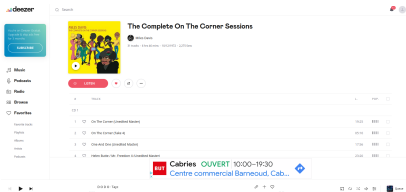
Voir aussi [une brève histoire du numérique](#), par Samuel Nowakowski.

Les transformations de la société par le numérique

Les "bons" aspects (*)...



WIKIPÉDIA
L'encyclopédie libre



(*) totalement subjectifs, bien entendu

Les transformations de la société par le numérique

... mais aussi des points plus problématiques

- L'influence du numérique sur la cognition (l'attention par exemple) [Citton, 2014, Desmurget, 2019]
- "Si c'est gratuit, c'est vous le produit !" :
 1. nos recherches sur Google nourrissent le moteur de recherche.
 2. Google revend nos données à des annonceurs (économie de la surveillance [Zuboff, 2019])
 3. nous sommes largement influencés par les résultats d'un algorithme de classement
- La plateformesation comme dérèglement de l'économie [Toledano, 2020]
- Les effets de polarisation de l'opinion sur les réseaux sociaux (bulles de filtre) ; les fake news etc

Les transformations de la société par le numérique

Diagnostic technique

- L'accélération des flux de données (l'ère du Zetaoctet)
- Le "pouvoir" des algorithmes à l'ère du "Big Data" (précision, prédiction)

Quelles conséquences ? [Cardon, 2018]

- Des algorithmes non neutres (trier ou choisir l'information)
- Loyauté (*trending topics*), effets inattendus (*Google suggest*), manipulation (*Google shopping*), déformation structurelle (*police prédictive*)
- Des réponses variées, de l'éducation à l'audit ou la régulation

Objectifs

- Décrire concrètement les interactions numérique/société via une introduction à l'analyse de réseaux sociaux (TP5)
- Comprendre et analyser un algorithme classique (page-rank) pour comprendre l'influence possible sur notre cognition ou nos comportements (TP6)
- Découvrir un mécanisme possible des bulles de filtre (TP7, sous réserve)

2 - Analyse de réseaux sociaux

Graphes (1)

Definition

En toute généralité un graphe $G = (V, E)$ est constitué d'un ensemble V de sommets ou noeuds et d'un ensemble $E \subset V \times V$ d'arêtes.

- Dans la suite, on supposera qu'il n'existe pas d'arêtes d'un sommet à lui-même (boucle). On parlera de *graphe simple*.
- Dans cette définition, on doit distinguer l'arête (a, b) de l'arête (b, a) . On parle alors de *graphe orienté*.
- Si l'ensemble des arêtes E est tel que $\forall (a, b) \in E, (b, a) \in E$, alors on parle de *graphe non orienté*.

Graphes (2)

Definition

La densité d'un graphe est le ratio du nombre d'arêtes d'un graphe sur le nombre d'arêtes possible

Pour un graphe dirigé et un graphe non dirigé, respectivement :

$$\frac{|E|}{|V|(|V| - 1)} \text{ et } \frac{|E|}{|V|(|V| - 1)/2}$$

Pouvez vous citer des exemples de graphes de la vie numérique, par exemple issus des réseaux sociaux, et pouvez vous les caractériser ?

Exemples

- Graphes non orientés : Facebook, LinkedIn,...
- Graphes orientés : Twitter, github,...

Plus de **détails** sur le nombre d'utilisateurs, la répartition géographiques etc.

Le nombre de Dunbar a été proposé pour borner le nombre de relations stables qu'un humain peut entretenir (148 dans l'étude d'origine).

Degrés (1)

Definition

Pour un graphe non orienté, on appelle voisinage d'un noeud i l'ensemble

$$\mathcal{V}(i) = \{j \in V; (i, j) \in E\}$$

Definition

Le degré $d(i)$ d'un noeud i est le nombre de voisin $d(i) := |\mathcal{V}(i)|$

On peut aussi définir les degrés *entrant* (resp. *sortant*) d'un noeud i pour un graphe orienté.

Degrés (2)

Definition

Le degré moyen d'un graphe non orienté est donné par la formule

$$\bar{d} = \frac{1}{|V|} \sum_{i=1}^{|V|} d(i)$$

Dans le cas d'un réseau social, on retrouve le nombre de Dunbar avec des estimations entre 100 et 200 [Gonçalves et al., 2011].

Proposition

On a $|V|\bar{d} = 2|E|$

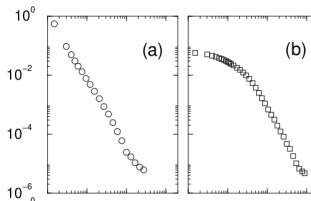
Le degré moyen n'est pas la statistique la plus intéressante qu'on puisse obtenir.

Degrés (3)

Definition

On peut définir la distribution des degrés comme étant la fonction $P : \mathbb{N} \rightarrow \mathbb{R}^+$ telle que $P(k)$ est la probabilité qu'un noeud soit de degré k . C'est à dire que :

$$P(k) = \frac{|\{i \in V; d(i) = k\}|}{|V|}$$



Internet (routeurs)

Graphe d'acteurs

Analyse des réseaux

Travaux fondateurs dans [Albert and Barabási, 2002]

Approche qui s'applique à différents types de réseaux, aussi variés que des réseaux trophiques ou le graphe des acteurs (Kevin Bacon's number)

Objectifs :

- Comprendre et décrire la structure de communication en analysant la topologie du graphe.
- Identifier et caractériser des sommets qui sont particulièrement importants.
- Comparer plusieurs graphes les uns avec les autres.

Métriques

On va considérer un certain nombre de métriques classiques qui concernent

- Les chemins du graphe ("mon nombre de Erdös est 5")
- La centralité des noeuds (rôle d'"influenceur")
- La transitivité ("deux de mes amis sont-ils amis?")

Chemins (1)

Definition

Un **chemin** de longueur p entre deux noeuds i et j est une suite finie de p arêtes (v_k, v_{k+1}) vérifiant $v_0 = i$ et $v_p = j$.

Definition

Un **plus court chemin** entre deux noeuds i et j est un chemin qui minimise la longueur des chemins entre i et j . On note $d(i, j)$ cette longueur.

Definition

La **longueur moyenne des chemins** est

$$\bar{L} = \frac{1}{n(n-1)/2} \sum_{i \neq j} d(i, j)$$

Chemins (2)

Comment calculer efficacement $\bar{L}$?

- Dijkstra ?
- Bellman-Ford ?
- Roy-Floyd-Warshall ! en $O(|V|^3)$

Algorithme de Roy-Floyd-Warshall

Data: Un graphe à n noeuds représenté par la matrice A telle que $A[i,j] = 1$ si (i,j) est une arête et $+\infty$ sinon.

Result: D matrice des distances paires à paires

$D = A$

```

for (k = 0 à n-1 ) do
    for (i = 0 à n-1 ) do
        for (j = 0 à n-1 ) do
            | D[i,j] = min( D[i,j], D[i,k]+D[k,j] )
        end
    end
end
end

```

Six degrés de séparation

Historique détaillé [ici](#).

- Karinthy, Milgram
- Expériences plus récentes
- Graphe des acteurs, des collaborations scientifiques, d'internet...

```

julienlefevre@Latitude-5480:~$ traceroute univ-amu.fr
traceroute to univ-amu.fr (139.124.244.54), 30 hops max, 60 byte packets
 1  livbox.hmc (192.168.1.1)  1.542 ms  4.481 ms  4.437 ms
 2  * * *
 3  ae104-0.hcnar101.rblc.orange.net (193.253.86.226)  13.364 ms  14.298 ms  14.780 ms
 4  ae61-0.nlnar101.rblc.orange.net (193.252.161.25)  15.749 ms  16.713 ms  18.075 ms
 5  ae60-0.nlnar102.rblc.orange.net (193.252.161.22)  19.364 ms  19.328 ms  20.456 ms
 6  193.252.137.54 (193.252.137.54)  28.934 ms  19.875 ms  19.951 ms
 7  renater-1-ge.opentransit.net (193.251.254.30)  22.417 ms  20.364 ms  18.834 ms
 8  et-2-0-2-ren-nr-parisi-rtr-131.noc.renater.fr (193.55.204.192)  19.562 ms et-3-1-1-ren-nr-parisi-rtr-131.noc.renater.fr (193.55.204.194)  21.892 ms et-2-0-2-ren-nr-parisi-rtr-131.noc.renater.fr (193.55.204.192)  19.271 ms
 9  * * *
10  et-2-0-0-ren-nr-lyon1-rtr-131.noc.renater.fr (193.51.180.107)  28.999 ms  29.104 ms  29.481 ms
11  xe-0-0-1-ren-nr-marseille-rtr-131.noc.renater.fr (193.51.177.17)  28.995 ms xe-0-0-17-ren-nr-marseille-rtr-131.noc.renater.fr (193.51.177.80)  28.438 ms xe-1-0-19-ren-nr-marseille-rtr-131.noc.renater.fr (193.55.204.110)  29.602 ms
12  * * *
13  ralnu-2-tel-0-marseille-rtr-021.noc.renater.fr (193.51.185.157)  27.510 ms  28.654 ms  28.168 ms
14  * * *
15  * * *
16  139.124.244.250 (139.124.244.250)  40.417 ms  37.032 ms  34.355 ms
17  betelgeuse2.pj.univ-amu.fr (139.124.244.54)  29.839 ms IX 34.238 ms IX 28.347 ms IX

```

Coefficient de clustering

Il s'agit de savoir à quel point les voisins d'un noeud sont connectés directement.

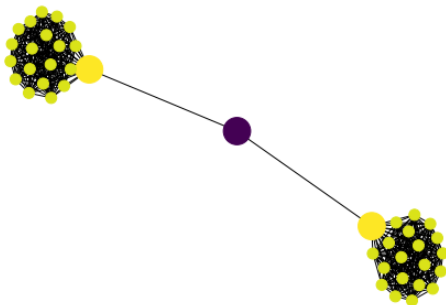
On définit le coefficient de clustering comme :

$$C(i) = \frac{\text{Nombre d'arêtes entre les voisins de } i}{\text{Nombre d'arêtes possibles entre les voisins}}$$

Le dénominateur s'exprime en fonction du degré $d_i(d_i - 1)/2$

Mesures de centralité (1)

Le degré d'un noeud donne une indication sur son importance relative mais présente aussi des limites.



Mesures de centralité (2)

Avec l'algorithme de page-rank on aura une façon élégante de formuler comment un noeud à une influence dans un graphe, en termes de *flux d'information* qui peut le traverser.

On a aussi des mesures de centralité basées sur la *structure des chemins*, dont le degré est un cas particulier (longueur 1). C'est le cas de la **centralité d'intermédiarité** (*betweenness centrality*) :

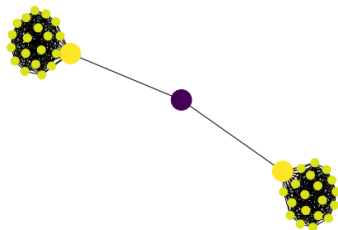
$$C_B(i) = \sum_{j \neq i, k \neq i} \frac{a_i(j, k)}{a(j, k)}$$

$a_i(j, k)$: nombre de plus courts chemins entre j et k passant par i

$a(j, k)$: nombre de plus courts chemins entre j et k

Mesures de centralité (3)

Cette mesure de centralité n'est pas forcément liée au degré comme on le voit en reliant deux graphes complets K_n .



$$C_B(n) = C_B(n+2) = (n-1)(n+1)$$

$$C_B(n+1) = n^2$$

$$C_B(i) = 0 \text{ sinon}$$

alors que

$$d(n) = d(n+2) = n$$

$$d(n+1) = 2$$

$$d(i) = n-1 \text{ sinon}$$

3 - Modèles de réseaux sociaux

Motivations

Il est intéressant de disposer de *modèles* pour générer des graphes dont les caractéristiques sont proches de ceux des graphes réels.

Ces modèles reposent sur des *hypotheses* qui peuvent avoir des interprétations concrètes et être plus ou moins adaptées à des situations réelles.

Dans la suite on disposera d'algorithmes pour générer des *graphes aléatoires*. On va voir qu'il existe des résultats bien établis sur les valeurs des métriques classiques qu'on peut obtenir.

Ces modèles ont été développés au départ dans le cadre de la théorie des graphes puis dans des communautés plus proches de la physique, en créant des liens interdisciplinaires (domaine des *systèmes complexes*).

Le modèle d'Erdős-Rényi

Il s'agit du premier modèle de graphe aléatoire introduit en 1959.

En entrée on considère :

- Le nombre de sommets n
- Le paramètre $p \in [0, 1]$ qui est la probabilité que deux sommets i et j pris au hasard soient reliés par une arête.

Pour générer de tels graphes, l'idée est de considérer chacune des $n(n-1)/2$ arêtes du graphe et de tirer indépendamment une variable aléatoire uniforme X sur $[0, 1]$. Si $X < p$ on garde l'arête dans le graphe.

Algorithmes

Data: n entier, $p \in [0, 1]$

Result: $G = (V, E)$ graphe

$V = \emptyset$, $E = \emptyset$

for ($i = 0$ à $n-1$) **do**

 Ajoute i à V

for ($j = i+1$ à $n-1$) **do**

X obtenu par tirage aléatoire uniforme sur $[0, 1]$

if $X < p$ **then**

 Ajoute (i, j) à E

end

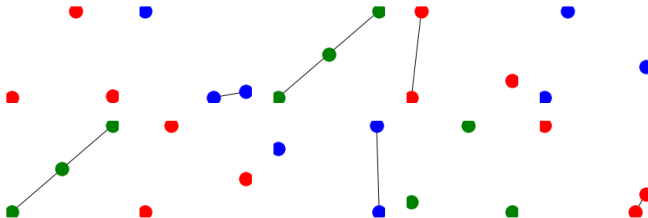
end

end

Exemples

Ce graphe est noté $G(n, p)$ et n'est pas unique puisqu'il est le résultat d'une expérience aléatoire.

Par exemple, pour $n = 3$ et $p = 1/3$, on peut obtenir



Propriétés des métriques

- Distribution des degrés

$$P(k) = C_n^k p^k (1 - p)^{n-k}$$

On peut calculer le degré moyen du graphe qui vaut $\bar{k} = np$

- Coefficient de clustering

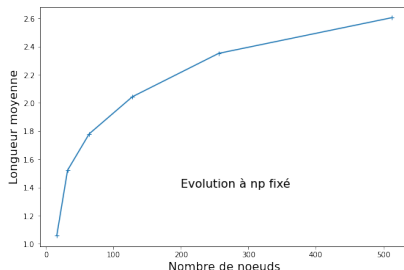
$$E(C_i) = p$$

- Longueur moyenne des chemins (plus difficile)

$$\bar{L} = O(\log n)$$

Propriétés des métriques

La longueur moyenne des chemins correspond à ce qu'on observe dans des graphes sociaux.

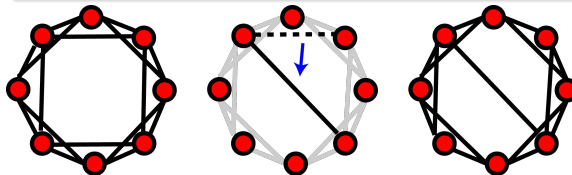


La distribution des degrés et le coefficient de clustering ne sont pas réalistes.

Vers d'autres modèles plus réalistes

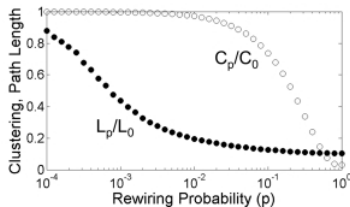
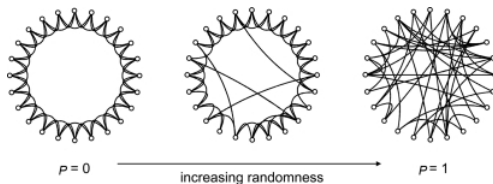
Modèle de Watts et Strogatz (1998)

On part d'un graphe en anneau où chaque sommet a $k/2$ voisins de chaque côté. Pour chaque sommet et pour chaque arête, il y a une probabilité p de recabler l'arête à un autre sommet au hasard.



Intuitivement on voit bien qu'on va créer des ponts dans le graphe (propriété "petit monde") tout en préservant les liens entre voisins (clustering élevé).

Vers d'autres modèles plus réalistes



En revanche, la densité des degrés suit le même type de loi que pour un graphe d'Erdős-Rényi.

Modèle de Barabasi-Albert

Ce modèle se base sur un principe d'*attachement préférentiel*.

- On part d'un graphe initial avec m_0 noeuds.
- A chaque étape on ajoute un noeud connecté à $m \leq m_0$ noeuds
- La probabilité de connexion d'un nouveau noeud i avec un noeud existant j dépend directement de d_j .

Algorithme

Data: $G_0 = (V_0, E_0)$ graphe avec $|V_0| = m_0$, $m \leq m_0$, n

Result: $G = (V, E)$ graphe avec $|V| = m_0 + n$

$V = V_0$, $E = E_0$

d_sum = Somme des degrés

for ($i = 0$ à $n-1$) **do**

 Ajoute $i + m_0$ à V

$curr_deg = 0$

while ($curr_deg < m$) **do**

X obtenu par tirage aléatoire uniforme sur $[0, 1]$

j noeud aléatoire différent de i et de ses voisins

if $X < \text{degré}(j)/d_sum$ **then**

 Ajoute (i, j) à E

$curr_deg += 1$, $d_sum += 1$

end

end

end

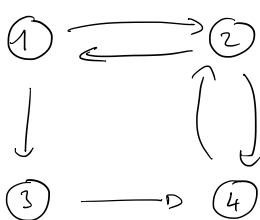
Propriétés des métriques

- Distribution des degrés de la forme $P(k) \sim k^{-3}$
- Coefficient de clustering en $n^{-3/4}$ mais avec une valeur 5 fois plus grande que pour les graphes d'Erdős-Renyi
- La longueur moyenne des chemins est en $\frac{\log n}{\log \log n}$

4 - PageRank

Matrice d'adjacence d'un graphe

Graphe = matrice :



target

source

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix}$$

Remarque : si le graphe est non orienté, on obtient une matrice symétrique ($\forall i, j, m_{i,j} = m_{j,i}$)

Matrice d'adjacence d'un graphe (2)

Les implémentations d'algos usuels (Dijkstra) utilisent la représentation en matrice d'adjacence.

Opérations de graphes = opération matricielle

Calcul des
successeurs :

$$\begin{array}{c} \text{source} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \end{array} \begin{pmatrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{pmatrix}$$

$$Y = X.M$$

$$\underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix}}_{\substack{\text{vecteur} \\ \text{représentant} \\ \text{le sommet } 1}} \underbrace{\begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}}_{\substack{\text{vecteur donnant} \\ \text{les sommets } s \text{ tels que} \\ 1 \rightarrow s}}$$

Matrices creuses (sparse matrix), algos adaptés

Matrice d'adjacence d'un graphe (2)

Les implémentations d'algos usuels (Dijkstra) utilisent la représentation en matrice d'adjacence.

Opérations de graphes = opération matricielle

Calcul des
successeurs :

$$\begin{array}{c} \text{source} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \end{array} \begin{pmatrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{pmatrix}$$

$$Y = X.M$$

$$\underbrace{(1 \ 0 \ 0 \ 0)}_{\substack{\text{vecteur} \\ \text{représentant} \\ \text{le sommet } 1}} \underbrace{\begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}}_{\substack{\text{vecteur donnant} \\ \text{les sommets } s \text{ tels que} \\ 1 \rightarrow s}}$$

Matrices creuses (sparse matrix), algos adaptés

Matrice d'adjacence d'un graphe (2)

Les implémentations d'algos usuels (Dijkstra) utilisent la représentation en matrice d'adjacence.

Opérations de graphes = opération matricielle

Calcul des
successeurs :

$$\begin{array}{c} \text{source} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \end{array} \begin{pmatrix} & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{pmatrix}$$

$$Y = X.M$$

$$\underbrace{(1 \ 0 \ 0 \ 0)}_{\substack{\text{vecteur} \\ \text{représentant} \\ \text{le sommet } 1}} \underbrace{\begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix}}_{\substack{\text{vecteur donnant} \\ \text{les sommets } s \text{ tels que} \\ 1 \rightarrow s}}$$

Matrices creuses (sparse matrix), algos adaptés

Un peu de dynamique

Epidémiologie : étude de la dynamique d'évolution d'une épidémie.

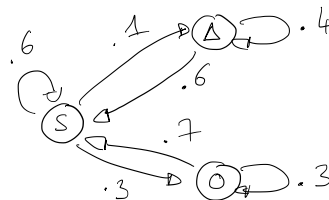
Modèle simple pour Covid 19,
deux variants, trois états :

- Sain
- Delta
- Omicron

Hypothèses stochastiques :

$\mathbb{P}(\text{Sain} \rightsquigarrow \text{Delta}) \approx 10\%$

$\mathbb{P}(\text{Sain} \rightsquigarrow \text{Omicron}) \approx 30\%$

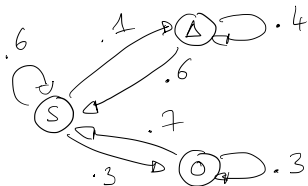


Comment estimer l'évolution de l'épidémie avec ce modèle ?

Chaînes de Markov

Propriété de Markov : aucune mémoire.

↪ L'évolution du système ne dépend que de son état courant, pas du passé.



$$\begin{array}{c}
 S \quad \Delta \quad O \\
 \begin{array}{l}
 S \\
 \Delta \\
 O
 \end{array}
 \begin{pmatrix}
 & S & \Delta & O \\
 \begin{pmatrix}
 .6 & .1 & .3 \\
 .6 & .4 & 0 \\
 .7 & 0 & .3
 \end{pmatrix}
 \end{pmatrix}
 \begin{array}{l}
 \Sigma = 1 \\
 \Sigma = 1 \\
 \Sigma = 1
 \end{array}
 \end{array}$$

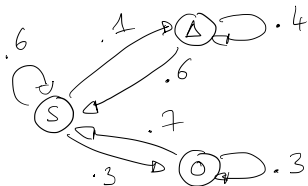
70% de sains, 30% de Delta : vecteur $X = (.7, .3, 0)$

Nouvelle distribution après une étape : $Y = X.M = (.6, .19, .21)$

Chaînes de Markov

Propriété de Markov : aucune mémoire.

↪ L'évolution du système ne dépend que de son état courant, pas du passé.



$$\begin{array}{c}
 S \quad \Delta \quad O \\
 \begin{array}{l}
 S \left(\begin{array}{ccc} .6 & .1 & .3 \\ .6 & .4 & 0 \\ .7 & 0 & .3 \end{array} \right) \\
 \Delta \\
 O
 \end{array}
 \begin{array}{l}
 \Sigma = 1 \\
 \Sigma = 1 \\
 \Sigma = 1
 \end{array}
 \end{array}$$

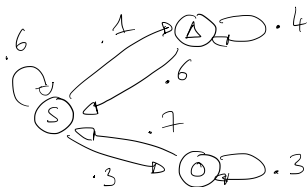
70% de sains, 30% de Delta : vecteur $X = (.7, .3, 0)$

Nouvelle distribution après une étape : $Y = X.M = (.6, .19, .21)$

Chaînes de Markov

Propriété de Markov : aucune mémoire.

↪ L'évolution du système ne dépend que de son état courant, pas du passé.



$$\begin{array}{c}
 S \quad \Delta \quad O \\
 \begin{array}{l}
 S \left(\begin{array}{ccc} .6 & .1 & .3 \\ .6 & .4 & 0 \\ .7 & 0 & .3 \end{array} \right) \\
 \Delta \\
 O
 \end{array}
 \begin{array}{l}
 \Sigma = 1 \\
 \Sigma = 1 \\
 \Sigma = 1
 \end{array}
 \end{array}$$

70% de sains, 30% de Delta : vecteur $X = (.7, .3, 0)$

Nouvelle distribution après une étape : $Y = X.M = (.6, .19, .21)$

Stabilisation

Definition

Une distribution de probabilité π est stationnaire si elle satisfait $\pi = \pi.P$.

Theorem

Sous certaines conditions, une chaîne de Markov possède une unique distribution stationnaire, vers laquelle la chaîne converge à partir de n'importe quelle distribution initiale.

Quelles conditions ?

- nombre fini d'états
- irréductible : il existe un chemin entre chaque paire de sommets
- apériodique : pgcd des longueurs des cycles = 1

Stabilisation

Definition

Une distribution de probabilité π est stationnaire si elle satisfait $\pi = \pi.P$.

Theorem

Sous certaines conditions, une chaîne de Markov possède une unique distribution stationnaire, vers laquelle la chaîne converge à partir de n'importe quelle distribution initiale.

Quelles conditions ?

- nombre fini d'états
- irréductible : il existe un chemin entre chaque paire de sommets
- apériodique : pgcd des longueurs des cycles = 1

Stabilisation

Definition

Une distribution de probabilité π est stationnaire si elle satisfait $\pi = \pi.P$.

Theorem

Sous certaines conditions, une chaîne de Markov possède une unique distribution stationnaire, vers laquelle la chaîne converge à partir de n'importe quelle distribution initiale.

Quelles conditions ?

- nombre fini d'états
- irréductible : il existe un chemin entre chaque paire de sommets
- apériodique : pgcd des longueurs des cycles = 1

Stabilisation (2)

Retour à l'exemple précédent.

Fait : la matrice précédente satisfait toutes les hypothèses.

Quelques exécutions :

[0.33, 0.33, 0.33]	[1, 0, 0]	[0.1, 0.7, 0.2]
[0.63, 0.17, 0.2]	[0.6, 0.1, 0.3]	[0.62, 0.29, 0.09]
[0.61, 0.13, 0.25]	[0.63, 0.1, 0.27]	[0.61, 0.18, 0.21]
[0.62, 0.11, 0.26]	[0.63, 0.1, 0.27]	[0.62, 0.13, 0.25]
[0.62, 0.11, 0.26]	[0.63, 0.1, 0.27]	[0.62, 0.11, 0.26]
[0.62, 0.1, 0.26]	[0.63, 0.1, 0.27]	[0.63, 0.11, 0.27]
[0.62, 0.1, 0.27]	[0.63, 0.1, 0.27]	[0.63, 0.11, 0.27]

Moteurs de recherche

Au coeur de la révolution numérique

Très gros profits : **exemple de Google** :

- profits doublés en 2021
- 76 Mds euros de bénéfice net

Chercher si une page web contient un mot-clé

Problème classique de recherche de motif dans une séquence.

Insuffisant : il faut tenir compte de l'importance des pages web

→ PageRank [Brin and Page, 1998]

→ Loyauté ? Robustesse ?

PageRank : présentation

Idee générale : web crawler = exploration aléatoire du web.

Voir le web comme une chaîne de Markov, et estimer sa distribution stationnaire.

Les noeuds les plus importants sont ceux de plus grande proba.

Algo réel tient compte de nombreux autres facteurs :

- pages officielles ont un coeff renforcé
- localisation géographique
- clics des utilisateurs
- ...

PageRank : présentation

Idee générale : web crawler = exploration aléatoire du web.

Voir le web comme une chaîne de Markov, et estimer sa distribution stationnaire.

Les noeuds les plus importants sont ceux de plus grande proba.

Algo réel tient compte de nombreux autres facteurs :

- pages officielles ont un coeff renforcé
- localisation géographique
- clics des utilisateurs
- ...

PageRank : présentation (2)

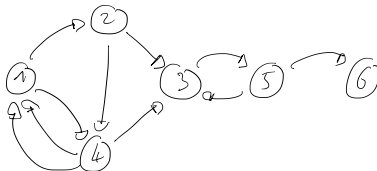
Modèle du surfer aléatoire : imaginons un agent qui se déplace aléatoirement dans le graphe du web. Quand il est sur une page, il choisit un lien sortant, avec proba uniforme sur l'ensemble des liens sortants.

En termes de graphe / matrice :

Graphe → Matrices d'adjacence qui compte le nombre de liens

→ (normalisation) : matrice contenant des probabilités

Exemple



$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$



$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 2/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}$$

Problème 1 : les puits

PageRank : les puits

Pour obtenir une chaîne de Markov irréductible, on donne la possibilité aux puits d'aller n'importe où, avec proba uniforme.

Exemple :

$$\begin{array}{c}
 \begin{array}{cccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
 \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \left(\begin{array}{cccccc}
 0 & 1/2 & 0 & 1/2 & 0 & 0 \\
 0 & 0 & 1/2 & 1/2 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 \\
 2/3 & 0 & 1/3 & 0 & 0 & 0 \\
 0 & 0 & 1/2 & 0 & 0 & 1/2 \\
 0 & 0 & 0 & 0 & 0 & 0
 \end{array} \right)
 \end{array}$$



$$\begin{array}{c}
 \begin{array}{cccccc}
 & 1 & 2 & 3 & 4 & 5 & 6 \\
 \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} & \left(\begin{array}{cccccc}
 0 & 1/2 & 0 & 1/2 & 0 & 0 \\
 0 & 0 & 1/2 & 1/2 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 \\
 2/3 & 0 & 1/3 & 0 & 0 & 0 \\
 0 & 0 & 1/2 & 0 & 0 & 1/2 \\
 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6
 \end{array} \right)
 \end{array}$$

PageRank : assurer la convergence

Problème 2 : assurer l'existence de la proba stationnaire.

Le graphe du web n'est pas fortement connexe : il existe plusieurs composantes connexes terminales.

Solution : ajouter une petite proba de reset : aller n'importe où.

$$P = \beta.M + (1 - \beta)J$$

où $J_{i,j} = 1/N, \forall i, j$ et N est la taille du web.

En pratique, on choisit $\beta \in [.85, .9]$

PageRank : assurer la convergence (2)

$$P = \beta.M + (1 - \beta)J$$

Avec cette nouvelle matrice on peut appliquer le théorème de convergence :

- nombre fini d'états
- irréductible (toutes les arêtes existent à cause de J)
- apériodique (boucle simple sur chaque état à cause de J)

PageRank : Calcul

Etant donné :

$$P = \beta.M + (1 - \beta)J$$

On veut calculer X tel que $X = X.P$

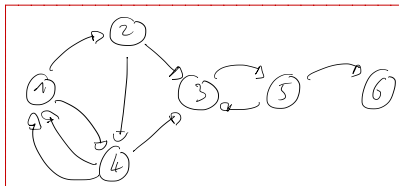
Remarque : X est le vecteur propre associé à la valeur propre 1.

Méthode des puissances :

calculer la suite (X_i) définie par $X_{i+1} = X_i.P$

- produit $X.P$: $O(N^2) \rightsquigarrow$ impossible ($n \approx$ trillion)
- M : matrice creuse : éviter ce calcul matriciel classique
- réécrire le produit $X.P$ en utilisant le fait que M est creuse

Exemple (suite)



→ $P =$

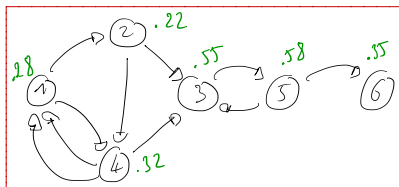
$$P = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 2/3 & 0 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 1/2 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix} \end{pmatrix}$$

↓

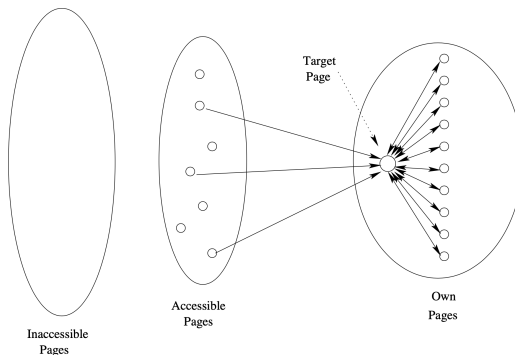
← $X = \begin{pmatrix} .28 \\ .22 \\ .55 \\ .32 \\ .58 \\ .35 \end{pmatrix}$

←

$P' = .9P + .1J$



Fermes de liens [Leskovec et al., 2014]

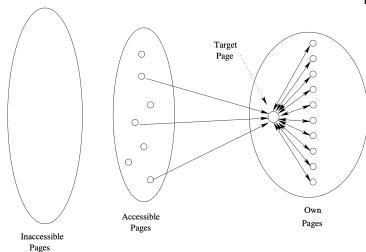


Objectif : améliorer le PageRank d'une page cible notée t (target)

Principe :

- être référencé par des "vraies" pages (forums...)
- amplifier son score grâce à des pages supplémentaires

Fermes de liens : analyse théorique



Notations :

- N : taille du web
- K : taille de la ferme
- y : valeur PageRank de t
- x : contribution PageRank des pages du web pointant vers t
- $P = \beta.M + (1 - \beta)J$

- PageRank d'une page de la ferme : $\beta \frac{y}{K} + \frac{1-\beta}{N}$
- PageRank de la page t :

$$y = x + \beta K \left(\beta \frac{y}{K} + \frac{1-\beta}{N} \right) + \frac{1-\beta}{N}$$

Fermes de liens : analyse théorique (2)

On en déduit :

$$y = \frac{x}{1 - \beta^2} + c \frac{K}{N}$$

avec $c = \frac{\beta}{1+\beta}$

Avec $\beta = 0.85$, on obtient :

- $\frac{1}{1-\beta^2} = 3.6$
- $c = 0.46$

➔ Détecter ce genre de techniques / Modifier l'algorithme pour les contrecarrer.

References I



Albert, R. and Barabási, A.-L. (2002).
Statistical mechanics of complex networks.
Reviews of modern physics, 74(1) :47.



Brin, S. and Page, L. (1998).
The anatomy of a large-scale hypertextual web search engine.
Comput. Networks, 30(1-7) :107–117.



Cardon, D. (2018).
Le pouvoir des algorithmes.
Pouvoirs, (1) :63–73.



Citton, Y. (2014).
Pour une écologie de l'attention.
Média Diffusion.



Desmurget, M. (2019).
La fabrique du crétin digital-Les dangers des écrans pour nos enfants.
Média Diffusion.



Gonçalves, B., Perra, N., and Vespignani, A. (2011).
Modeling users' activity on twitter networks : Validation of dunbar's number.
PloS one, 6(8) :e22656.



Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014).
Mining of Massive Datasets, 2nd Ed.
Cambridge University Press.

References II



[Toledano, J. \(2020\).](#)

GAFA : reprenons le pouvoir !

[Odile Jacob.](#)



[Zuboff, S. \(2019\).](#)

The age of surveillance capitalism : The fight for a human future at the new frontier of power :

Barack Obama's books of 2019.

[Profile books.](#)