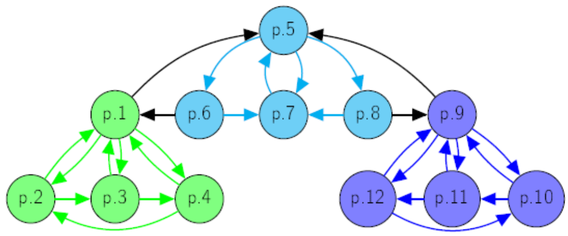
Le mardi 11 Avril 2023, les élèves de première suivant la Spécialité NSI ont assisté à un exposé intitulé « L'algorithme PageRank de Google »

Cet exposé a été effectué par **Frédéric Paccaut, maitre de conférences à l'Université Jules Verne de Picardie et chercheur au Laboratoire Amiénois de Mathématiques Fondamentales et Appliquées**.

L'objectif de l'exposé de Frédéric Paccaut est de nous **expliquer le principe de l'algorithme du système de classement des pages Web, le PageRank, utilisé par le moteur de recherche Google**.

Le web est l'ensemble des pages reliées entre elles par des liens hypertextes, sur Internet. Il correspond à un objet mathématique qui s'appelle un graphe : les pages sont les sommets de ce graphe et les liens hypertextes entre les pages sont les arêtes du graphe.
Pour faciliter la compréhension du fonctionnement du web, Frédéric Paccaut a pris un exemple avec 12 pages. Les flèches, arêtes du graphe, permettent de connaitre quelle page est accessible à partir de quelle autre page.

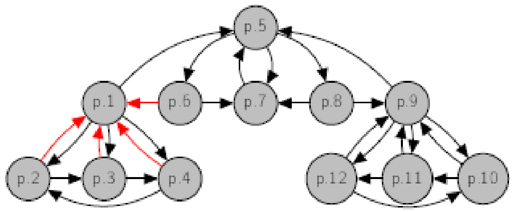


Lorsqu'on cherche par exemple "google" sur le moteur de recherche Google, on trouver 25 270 000 000 résultats (en 0,31 secondes !). Ces résultats sont donnés dans un ordre précis, qui se veut pertinent.
Mais comment faire pour classer par ordre d'importance les pages obtenues ? Le seul moyen de classer ces résultats est d'utiliser les liens des pages entre elles dans le but d'associer à chaque page un rang.

Reprenons le graphe de 12 pages donné en exemple. Il se dégage dans ce graphe trois groupes assez indépendants. Les pages 1 et 9 semblent être les pages importantes de deux sous-graphes car de nombreuses pages pointent vers elles. On peut s'attendre à ce que la page 5 soit alors la plus importante du graphe car les pages 1 et 9 pointent vers elle.

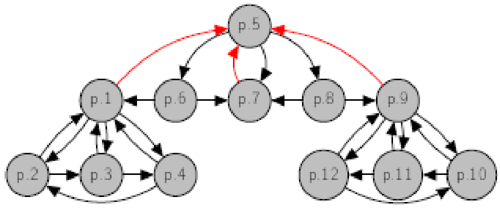
Mais comment chiffrer cette intuition pour pouvoir traiter des graphes plus volumineux ?

■ On peut dans un premier temps tout simplement compter le nombre de liens qui pointent sur chaque page. On désigne alors par s_i le nombre de liens pointant vers la page i.



On obtient alors

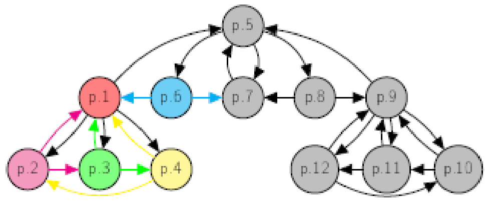
$$(s_1,\ s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}) = (4,\ 2,\ 2,\ 2,\ 3,\ 1,\ 3,\ 1,\ 4,\ 2,\ 2,\ 2)$$



Le classement obtenu ne correspond pas à notre intuition, qui définit la page 5 comme étant la plus importante.

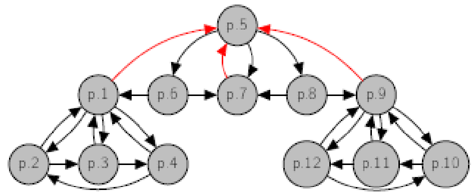
■ L'idée pourrait donc être de ne pas accorder la même importance à tous les pointeurs. On obtiendrait alors un classement qu'on peut alors qualifier de pondéré.
Comment pondérer ? La page 2 pointe vers deux pages différentes donc à partir de la page 2, il y a une chance sur 2 pour qu'on se dirige sur l'une des deux pages accessibles à partir de la page2 ; chaque lien sortant de la page 2 sera donc pondéré par ½. De même, si une page pointe vers trois pages différentes, chaque lien comptera pour 1/3...

Alors $s_1 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}$ $\qquad s_5 = \frac{1}{4} + 1 + \frac{1}{4}$



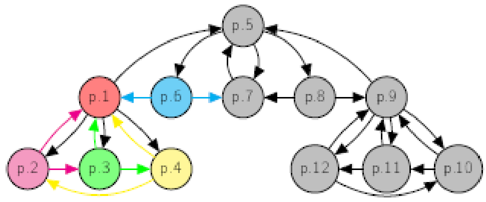
On obtient alors :

$$(s_1,\ s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}) = \left(2,\ \frac{3}{4},\ \frac{3}{4},\ \frac{3}{4},\ \frac{3}{2},\ \frac{1}{3},\ 3,\ 1,\ 4,\ 2,\ 2,\ 2\right)$$



Encore une fois, le résultat obtenu ne rend pas compte de notre ressenti car la page 5 n'apparait toujours pas comme étant la plus importante.

■ On va alors utiliser ce qu'on appelle un classement récursif, en tenant compte de l'importance de chaque page et en la pondérant comme expliqué précédemment.



On obtient alors les deux égalités suivantes pour les pages 1 et 5 :

$$s_1 = \frac{1}{2}\, s_2 + \frac{1}{2} s_3 + \frac{1}{2} s_4 + \frac{1}{2} s_6 \qquad s_5 = \frac{1}{4} s_1 + s_7 + \frac{1}{4} s_9$$

On peut alors écrire 12 égalités, une par page, et dans chaque égalité peuvent intervenir les autres pages : on a un système linéaire de 12 équations à 12 inconnues.

$$\left\{\begin{array}{lcl}
s_1 & = & \frac{1}{2}s_2 + \frac{1}{2}s_3 + \frac{1}{2}s_4 + \frac{1}{2}s_6 \\
s_2 & = & \frac{1}{4}s_1 + \frac{1}{2}s_4 \\
s_3 & = & \frac{1}{4}s_1 + \frac{1}{2}s_2 \\
s_4 & = & \frac{1}{4}s_1 + \frac{1}{2}s_3 \\
s_5 & = & \frac{1}{4}s_1 + s_7 + \frac{1}{4}s_9 \\
s_6 & = & \frac{1}{3}s_5 \\
s_7 & = & \frac{1}{3}s_5 + \frac{1}{2}s_6 + \frac{1}{2}s_8 \\
s_8 & = & \frac{1}{3}s_5 \\
s_9 & = & \frac{1}{2}s_8 + \frac{1}{2}s_{10} + \frac{1}{2}s_{11} + \frac{1}{2}s_{12} \\
s_{10} & = & \frac{1}{4}s_9 + \frac{1}{2}s_{12} \\
s_{11} & = & \frac{1}{4}s_9 + \frac{1}{2}s_{10} \\
s_{12} & = & \frac{1}{4}s_9 + \frac{1}{2}s_{11}
\end{array}\right.$$

Ce système, pénible à résoudre à la main, se résout bien avec un ordinateur et on obtient une solution qui semble plus cohérente avec notre ressenti initial car la page 5 est alors bien identifiée comme étant la plus importante :

$$(s_1,\ s_2, s_3, s_4, s_5, s_6, s_7, s_8, s_9, s_{10}, s_{11}, s_{12}) = (2,\ 1,\ 1,\ 1,\ 3,\ 1,\ 2,\ 1,\ 2,\ 1,\ 1,\ 1)$$

Dans la pratique, il n'y a pas que les liens de 12 pages à étudier : ce sont plusieurs millions d'équations à plusieurs millions d'inconnues qui sont à résoudre et des méthodes spécialisées sont nécessaires. Le calcul matriciel entre alors en scène et la théorie de Perron-Frobenius permet de trouver les scores de chaque page.

C'est grâce à divers domaines mathématiques que Google fonctionne aussi bien et que notre quotidien d'internaute est si agréable et enrichissant.
Sans les mathématiques, nous serions encore à devoir nous déplacer dans les bibliothèques et à y errer de longues heures à la recherche des informations que nous cherchons alors qu'avec Internet, quelques clics suffisent pour les obtenir !