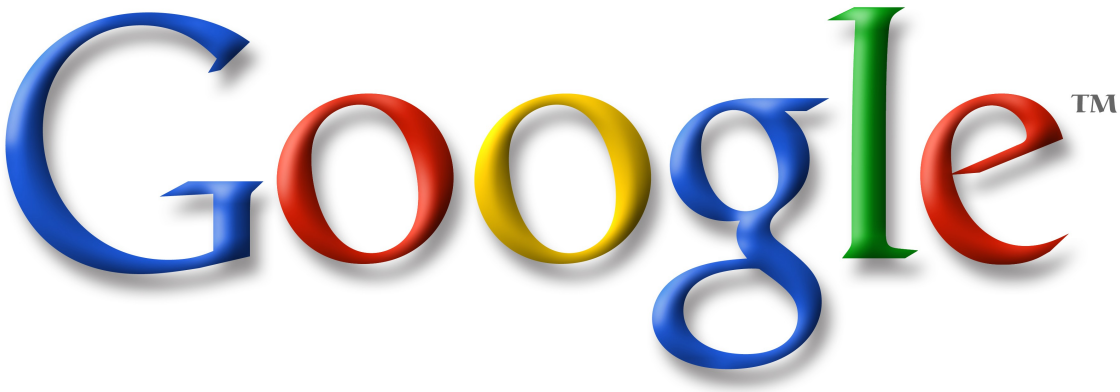
Par Gabriel LELLOUCH, Lucas TORRENS et Hugo VIOULAC




# Le PageRank
# Modélisation matricielle du fonctionnement

Edité par *Les amis des mathématiques.*

Pour Alain Combrouze,
de la part de Titi, Toto et les autres

« Les propositions mathématiques sont reconnues comme vraies
parce que personne n'a intérêt qu'elles soient fausses. »
(Montesquieu)

« Imaginez le groupe monogène, à la fin du film : « tu es fini ! » ;
le film se finit en concluant que le groupe est cyclique. »
(A. Combrouze)

# LE PAGERANK
# MODÉLISATION MATRICIELLE DU FONCTIONNEMENT

LELLOUCH GABRIEL, TORRENS LUCAS ET VIOULAC HUGO
MPSI V

Table des matières

## 1. Introduction

Depuis quelques décennies, Internet est devenu un outil très utilisé. De nombreux sites se sont développés, et désormais, on peut tout faire par Internet, encore faut-il savoir où aller, à qui s'adresser , avoir une idée de la fiabilité d'un site. A la manière d'une araignée, Google a tissé sa toile sur le Web depuis dix ans, recensant et classant près de 1000 milliards de pages web (rien que ça...). Google a écrasé la concurrence et désormais ne la craint plus. Les premiers moteurs (Altavista ou Yahoo) ne faisaient qu'indexer, c'est à dire trouver toutes les pages contenant le ou les mots clés recherchés. On pouvait retrouver les pages contenant un mot clé donné mais les résultats n'étaient pas triés de manière efficace. Le nombre de fois où apparaissait le mot clé faisait apparaître la page en haut de la liste de résultat, ce qui n'est pas vraiment pertinent. Le nombre de répétions du mot clé n'est pas un critère intéressant puisque cette donnée est aisément falsifiable. Il faut faire une analyse plus fine du Web pour être capable de mesurer automatiquement l'importance de chaque site. Sergey Brin et Lawrence Page, étudiants à Standford ont trouvé une solution aussi originale que simple : utiliser l'information des liens entre les pages pour mesurer l'importance des sites, et être alors capable de classer correctement les résultats d'une recherche de mots clés. Ils ont alors, et ceci sans la moindre publicité, très vite pris le dessus sur leurs concurrents directs.

Aujourd'hui, Google est avant tout une société fondée le 27 septembre 1998 dans la Silicon Valley, en Californie, par Larry Page et Sergey Brin, auteurs du moteur de recherche Google. Google posséderait le parc de serveurs le plus important du monde, avec environ 500 000 machines réparties sur 32 sites à travers le monde. Il s'agit aujourd'hui d'une véritable entreprise qui s'est organisée autour d'un projet étonnant, celui de répertorier puis de classer par ordre de pertinence les pages du Web et ainsi proposer un service optimum de choix pour les internautes, quel que soit leur centre d'intérêt.

Nous connaissons tous le principe, dans nos favoris ou par l'intermédiaire de la barre de tâche de Google, on mène une recherche grâce à Google afin d'être très vite dirigé vers ce que l'on cherche, et que l'on sait pouvoir trouver. Ceci est devenu depuis peu le quotidien de chacun, les fameux mots clés nous permettant en un clic de consulter une page qui bien souvent correspond idéalement à notre demande.

Mais alors comment cela fonctionne, comment est-ce qu'en un quart de seconde, les serveurs de Google sont capables de renvoyer exactement les pages les plus intéressantes ? Comment est stockée cette quantité incroyable d'informations ?

Les petits secrets de cette entreprise sont bien évidemment confidentiels autant que l'est la recette du bon vieux Coca-Cola© ; cependant, l'algorithme matriciel du PageRank est dévoilé sous une forme quelque peu simplifiée et accessible aux mathématiciens intéressés.
Cet ouvrage est donc destiné à présenter son fonctionnement. Nous aborderons tout d'abord l'idée initiale, celle d'un vecteur de probabilité invariant, qui représenterait les N pages (de l'ordre de $10^{10}$) du Web pondérées par leur indice de pertinence. Ensuite de quoi nous verrons comment est construite la matrice Google, stochastique, et quelles sont ses principales caractéristiques. Nous nous intéressons après cela aux paramètres qui influent sur la pertinence et la vitesse de convergence du vecteur invariant. Puis, on présentera un modèle matriciel traduisant le cas d'un Web-jouet d'une dizaine de pages.

## 2. Principe général du PageRank

Avant toute chose, il est nécessaire, pour comprendre le fonctionnement de Google, de comprendre le principe général de la méthode de classement des pages. Nous vous proposons dans cette première partie un rapide tour d'horizon des moyens mis en jeu dans la mise en place de la hiérarchie des sites Internet, en commençant par introduire les éléments de base de cette méthode qui aujourd'hui a permis à ses inventeurs de se hisser au rang de concepteurs du plus grand moteur de recherche planétaire.

### 2.1. **Représenter le Web sous forme d'un graphe.**

Commençons tout d'abord par donner une définition : Le PageRank ou PR est le système de classement des pages Web utilisé par le moteur de recherche Google pour déterminer l'ordre et la pertinence des liens dans les résultats de recherche qu'il fournit, autrement dit, pour classer par ordre d'importance les différentes pages. De nos jours le PageRank n'est plus l'unique indice entrant en jeu dans l'algorithme qui permet de classer les pages Internet dans les résultats de recherche de Google, mais il est celui qui est la base de son originalité. Ce système a été inventé par Larry Page, cofondateur de Google. Ce mot est une marque déposée.

Le PageRank est le résultat de scores que Google attribue aux pages du Web essentiellement en fonction de leur popularité, c'est-à-dire du nombre de pages qui "pointent" vers cette même page, autrement dit le nombre de pages qui présentent un lien hypertexte permettant par un simple clic de se trouver redirigé vers le site correspondant. En fait, le principe général du PageRank est le suivant : une page sera d'autant plus importante que des pages importantes pointeront vers elle. Nous verrons par la suite comment ce principe, apparemment abstrait, peut être mis en œuvre.

Quel serait alors l'outil mathématique adapté aux classements des pages ? L'outil vectoriel semble le plus efficace. Le but du PageRank n'est autre que d'établir un vecteur ligne, dont la taille correspond au nombre de pages Web, et dont chaque coefficient indique la pondération, le degré de pertinence d'une page Web. Ce vecteur sera d'ailleurs appelé vecteur PageRank, et on parlera, pour qualifier ce coefficient propre à une page, du PageRank de cette page. On rappelle l'ordre

de grandeur du nombre de pages Web, qui est d'environ $10^{10}$, ce qui montre à quel point la pondération est complexe et difficile à stocker. Le PageRank s'appuie donc sur la création d'un véritable graphe du Web dans lequel sont mis en jeu la totalité des pages dans le monde entier. Ce graphe, qui a autant de nœuds qu'il existe de pages Web, indique les liens hypertextes à la surface de la Toile, ce qui est à la base du calcul du PageRank de chaque page.

Ainsi, la première étape pour avoir une idée générale du PageRank consiste à se représenter ce graphe du Web. En effet, il est alors possible d'avoir une idée, certes extrêmement imprécise, mais assez générale du PageRank de chaque page, par son importance dans ce graphe. Plus nombreuses seront les pages qui pointeront vers elle, plus cette page aura de chances d'avoir un PageRank élevé. Bien sûr, plus les pages qui pointent vers elle seront importantes, plus sa pertinence sera considérée comme élevée. C'est pourquoi il faut d'abord déterminer le PageRank des pages pointant vers cette page-là, en recommençant le même processus indéfiniment. Mais l'important semble donc de calculer le PageRank de chaque page, et ce vecteur PageRank a en fait une propriété remarquable : il s'inscrit dans le cadre d'une chaîne de Markov.

## 2.2. **Un vecteur au sein d'une chaîne de Markov.**

On appelle chaîne de Markov une suite de variables aléatoires $(X_n)$ telle que, pour chaque $n$, connaissant la valeur de $X_n$, $X_{n+1}$ soit indépendante de $X_k$, pour $k$ inférieur ou égal à $n-1$.

Autrement dit, pour tout $n$ et pour toutes valeurs possibles $i_0, ..., i_n, i_{n+1}$, la probabilité que $X_{n+1}$ prenne la valeur $i_{n+1}$ sachant que $X_0 = i_0, X_1 = i_1, ..., X_{n-1} = i_{n-1}$ et $X_n = i_n$ ne dépend que de $i_{n+1}$ et de $i_n$.

C'est pourquoi le vecteur PageRank s'inscrit effectivement dans une chaîne de Markov. En effet, si un surfer fait une recherche qui le dirige vers un site et que celui-ci lui propose des liens vers d'autres sites, la probabilité que ce surfer se dirige vers un autre site après est affectée logiquement par son opération précédente, c'est à dire sa présence sur le site, et ainsi de suite. Les chaînes de Markov interviennent donc dans les probabilités de présence sur les pages Web, et c'est cela qui est crucial. Le vecteur PageRank prend donc en compte les opérations des

internautes pour s'adapter à la demande en direct. Ainsi, le nombre d'opérations effectuées affecte notre PageRank.

En fait, ce modèle d'un surfer à la surface du Web est à la base de tout raisonnement sur le PageRank. Avec cette notion de chaîne de Markov, si l'on considère un individu lambda qui surfe sur le Web de manière aléatoire, le vecteur PageRank doit représenter la probabilité qu'a ce surfeur de se diriger vers telle ou telle page. Chacun de ses $10^{10}$ coefficients correspond à la chance de se retrouver sur une page pour un surfer aléatoire. Cela rejoint l'idée première de classer les sites par importance, car plus une page est importante, plus on a de chance de s'y retrouver. Et cette interprétation donne à notre vecteur de nombreuses propriétés.

Tout d'abord, le vecteur PageRank est un vecteur qui doit représenter la probabilité qu'un internaute soit intéressé par une page et donc s'y dirige. Les coefficients sont donc strictement positifs, puisque la probabilité d'accès à une page ne peut pas être inférieure ni égale à 0 (étant donné que l'on part du principe que toute page est accessible et qu'il n'existe pas de pages qui ne sont jamais visitées, ce qui paraît logique). Un indice de probabilité étant nécessairement compris entre 0 et 1 (ici strictement), les coefficients du vecteur PageRank sont donc tous dans l'intervalle ]0,1[.

Une autre propriété intéressante relative au PageRank, qui découle de son interprétation comme indice de probabilité, concerne la somme de ses coefficients. En effet, ce vecteur représente la probabilité d'accéder à une page précise. Ainsi la probabilité d'accéder à une page quelconque quand on va sur Internet est la probabilité qui correspond à une certitude : elle vaut 1. De plus, la probabilité d'accéder à une page quelconque correspond à la somme des coefficients du vecteur PageRank. La somme des PageRanks de toutes les pages est donc égale à 1.

Ainsi, le PageRank s'inscrit dans une chaîne de Markov et possède de ce fait les caractéristiques d'un vecteur de probabilité, ce qui sera très utile par la suite, dans son calcul et son utilisation. Sa construction est donc pertinente au vu de son usage. Mais il est maintenant temps d'introduire l'autre base du fonctionnement de Google, qui, associée au PageRank, permettra une utilisation simple et rapide : La matrice Google.

### 2.3. **L'obtention d'un vecteur invariant.**

La matrice Google, ou matrice $G$, dont la construction sera détaillée dans le chapitre suivant, est une matrice carrée dont la taille correspond au nombre de pages recensées sur le Web. Son but est de représenter les liens entre pages grâce à des coefficients numériques. En effet, s'il est trop difficile de traiter un véritable graphe de $10^{10}$ pages, il peut paraître logique de rassembler les informations de ce graphe dans une matrice, outil privilégié en informatique. Le coefficient de la ligne $i$ et de la colonne $j$ correspondra en fait à la présence et à l'importance d'un lien de la page $i$ pointant vers la page $j$. On sent ainsi que l'on n'est pas loin du vecteur de probabilité construit précédemment, qui doit représenter l'importance d'une page à travers les liens qui la joignent aux autres pages. C'est en fait cette matrice Google qui permettra de calculer le vecteur PageRank.

En effet, l'idée générale est d'obtenir un vecteur invariant par la multiplication à droite par la matrice Google. Cela semble tout à fait logique car si la matrice Google représente le graphe du Web et les liens entre les pages, le produit par un vecteur-ligne à gauche va affecter ce dernier en modifiant le $j$-ième coefficient en fonction de la $j$-ième colonne de la matrice $G$, c'est-à-dire en fonction de l'importance des liens pointant vers la page $j$. Si l'on arrive à trouver un vecteur invariant, de probabilité, il serait tout à fait relié à la structure du Web, et correspondrait en fait à notre fameux vecteur PageRank. Le calcul utilisé par Google dans la détermination de ce vecteur est détaillé dans le chapitre 6.

Cette matrice Google se doit cependant d'avoir un certain nombre de caractéristiques nécessaires au bon fonctionnement et au calcul du PageRank. En effet, les résultats connus sur les chaînes de Markov, ainsi que sur la théorie de Perron-Frobenius, qui est au cœur du principe du PageRank, exigent par exemple la stochasticité de la matrice $G$, pour ne serait-ce qu'avoir l'existence de ce vecteur invariant. Il est bien sûr évident que Page et Brin ont tenu compte de ces caractéristiques dans la construction de la matrice.

Le principe général de la méthode de classement des pages du Web est donc assez simple, et tient en deux éléments : un vecteur et une matrice. Le vecteur PageRank, de taille $1 \times N$, où $N$ est le nombre de pages Web, contient des coefficients correspondant à l'indice de popularité de chaque page. Il se doit d'être de probabilité (c'est-à-dire à coefficients positifs et de somme 1), et nous verrons en partie 4 qu'il

existe et est unique, ce qui est primordial dans son but. La matrice $G$ Google, de taille $N \times N$ représente, elle, un véritable graphe du Web en donnant les relations de liaisons entre les différentes pages. Elle se doit de rendre le vecteur PageRank invariant par multiplication, ce qui permettrait un classement efficace et réaliste des pages Web par ordre de pertinence. Le vecteur et la matrice Google s'inscrivent dans une chaîne de Markov afin de correspondre aux besoins en temps réel des utilisateurs de Google.

Mais nous n'avons que trop parlé de théorie : place à présent à la construction de Google, avec comme première étape, la formation de la matrice Google !

## 3. Construction de la matrice $G$ de Google

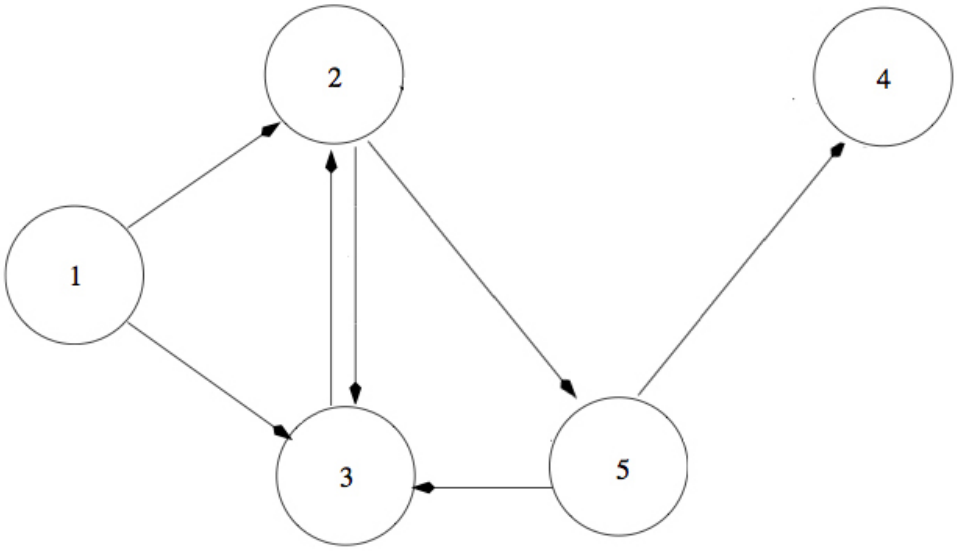
Le PageRank, cette donnée indispensable à cette entreprise titanesque que constitue la classification des milliards de pages engrangées par Internet, sera ici représenté sous la forme d'un vecteur-ligne, et noté $\pi$.

Le calcul du PageRank se fait à l'aide de la méthode de la puissance (se référer à la partie du même nom pour de plus amples détails), en itérant le calcul suivant :

$$\pi^{(k+1)} = \pi^{(k)} G$$

où $\pi^{(k)}$ représente le vecteur PageRank obtenu lors de la k-ième itération, et $G$ une matrice de type $N \times N$, nommée matrice Google, qui dépend de la structure même du Web.

L'objectif de cette partie sera de voir quelles furent les différentes étapes de la construction de cette matrice $G$, et nous utiliserons, pour illustrer le principe de la méthode, le mini-web suivant (un exemple plus conséquent sera traité plus tard) :



### 3.1. **Une matrice préliminaire.**

Une première idée (certes simpliste, mais il faut bien commencer d'une manière ou d'une autre) consisterait à construire une matrice $L = (l_{ij})$ telle que $l_{ij} = 1$ si la $i$-ème page présente un lien menant vers la $j$-ème page, 0 sinon. Pour l'exemple précédent, on obtient :

$$L = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

Il est toutefois de notoriété publique, ou presque, que les matrices stochastiques présentent des propriétés intéressantes dans le cadre des chaînes de Markov, on modifie donc légèrement l'idée initiale, et l'on définit une matrice $H = (h_{ij})$ telle que $h_{ij} = \dfrac{1}{l_i}$, où $l_i$ représente le nombre de liens présents sur la page $i$, si $i$ pointe vers $j$, 0 sinon. Cela revient à diviser les éléments de la ligne $i$ de $L$ par le nombre d'élément de cette même ligne. Pour le mini-web précédent, on obtient :

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

Une conséquence immédiate de cette construction, mise à part, bien sûr, la stochasticité des lignes non nulles de la matrice $H$, réside dans le fait qu'il s'agit d'une matrice très creuse, avantage non négligeable si l'on veut opérer sur des matrices d'une telle envergure (n'oublions pas qu'il s'agit là de matrices d'ordre $10^{10}$).

On peut effectivement considérer qu'en moyenne, une page web renvoie vers dix autres pages, donc H aurait environ $10N$ éléments non nuls, soit $10^{11}$, alors qu'elle contient $10^{20}$ éléments. Les calculs menés passent donc d'une complexité quadratique à une complexité linéaire.

Ne nous laissons cependant pas leurrer par ces quelques avantages, car une telle construction n'est pas sans générer quelques problèmes.

De fait, il existe sur la Toile nombre d'éléments ne renvoyant vers aucune autre page (notamment des images, des sons ou des fichiers .pdf). On peut prendre pour exemple la page portant le numéro quatre du petit graphe initial. Un autre défaut de cette structure est engendré par les cycles : ces éléments dérangeants sont constitués de pages qui pointent successivement vers d'autres pages, de telle sorte que l'on revient finalement vers la page de départ. L'exemple le plus simple serait un couple de pages dont l'une pointe exclusivement vers l'autre, et réciproquement. En quoi sont-ce des problèmes ?

Considérons en effet un internaute aléatoire, autrement dit un individu qui choisirait toujours au hasard un lien parmi ceux proposés par la page sur laquelle il se trouve. Ainsi, si ce surfer passe beaucoup de temps sur une même page $i$, on peut en déduire que nombreuses sont les pages $j$ qui pointent vers $i$.

Les problèmes cités plus tôt apparaissent alors évidents : en effet, un tel internaute, s'il en vient à être dirigé sur un fichier sans liens, ne peut clairement plus continuer sa promenade virtuelle. D'un autre côté, s'il se retrouve sur un cycle, autrement dit un ensemble fini de pages se pointant l'une l'autre de telle manière que l'on revienne sans cesse sur ces mêmes pages, il ne peut rien faire d'autre que se mettre à tourner en rond, visitant toujours les mêmes pages. Voyons donc quelle fut la réponse de Brin et Page à ce problème.

### 3.2. **Un ajustement de stochasticité.**

La résolution de ladite difficulté aurait certes pu se faire grâce à la théorie des chaînes de Markov, eussions-nous considéré $H$ comme une matrice de probabilité de transition (autrement dit, une matrice stochastique dont tous les coefficients sont positifs), mais il apparaît bien plus simple de compléter le comportement de notre internaute aléatoire : celui-ci, s'il se retrouve sans possibilité de choisir un lien et ainsi continuer sa marche aléatoire, aura alors la possibilité de se rendre sur n'importe quelle page du Web, et ce avec une parfaite équiprobabilité. Ceci revient en fait à remplacer les lignes de zéros de H par des lignes de $\frac{1}{N}$. On peut alors définir une nouvelle matrice $S$ :

$$S = H + \mathrm{a} \cdot \frac{1}{N} {}^t e$$

où $a_i = 1$ si la page $i$ est un fichier sans liens, 0 sinon ; $e$ est quant à lui le vecteur colonne ne contenant que des 1 ; a est un vecteur-colonne. Pour le schéma précédent, la matrice $S$ est la suivante :

$$S = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1 & 0 & 0 & 0 \\ 1/5 & 1/5 & 1/5 & 1/5 & 1/5 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$

On obtient alors une nouvelle matrice $S$ entièrement stochastique en ligne. Ce dernier élément, s'il corrige en partie les défauts de la précédente structure, n'est néanmoins pas suffisant pour assurer à lui seul la convergence du vecteur invariant $\pi$. Nous verrons plus tard qu'il est en fait nécessaire pour cela que tous les coefficients de notre matrice stochastique Google doivent être strictement positifs, et plus encore, la matrice doit avoir une unique valeur propre de module maximal.

### 3.3. **La matrice $G$ de Google.**

Ce dernier ajustement peut toujours être justifié par notre ami aléatoire, cette simulation devenue indispensable au bon déroulement des opérations. Nous nous sommes jusqu'ici contenté de prendre en compte les liens existant entre les différentes pages d'Internet, sans prendre en compte aucun autre facteur de navigation.

Il n'est pas rare en effet qu'un internaute lassé de ses pérégrinations virtuelles abadonne purement et simplement le site où il se trouvait pour se rendre sur un autre site, sans aucun lien avec le précédent. Et c'est là que nous définissons la véritable et unique, l'inimitable et innovante matrice Google, de la manière suivante :

$$G = \alpha S + (1 - \alpha)\tfrac{1}{N}E$$

où $E = e \cdot {}^t e$, matrice de type $N \times N$ ne contenant que des 1, sera nommée matrice de téléportation (on pourra utiliser, plus généralement, un vecteur de probabilité quelconque $v$, et définir $E = e \cdot {}^t v$). $\alpha$ est quant à lui un paramètre réel, choisi entre 0 et 1.

Ce paramètre $\alpha$, dont nous constaterons l'importance plus que cruciale ultérieurement, est intimement lié à la probabilité qu'a notre surfer de se “téléporter" de la manière décrite plus haut. Ainsi, si $\alpha$ vaut 0,4, il

faudra comprendre qu'un internaute lambda a 60% de chances de se téléporter, contre 40% de suivre simplement la structure prédéfinie de la Toile. La matrice $E$ étant clairement uniforme, on comprend que ce phénomène de téléportation est équiprobable (compte tenu de toutes les pages du web). Pour le petit exemple que nous avions préalablement choisi, nous obtenons (avec $\alpha = 0,85$) :

$$G = \begin{pmatrix} 0,030 & 0,45 & 0,45 & 0,030 & 0,030 \\ 0,030 & 0,030 & 0,45 & 0,030 & 0,45 \\ 0,030 & 0,88 & 0,030 & 0,030 & 0,030 \\ 0,2 & 0,2 & 0,2 & 0,2 & 0,2 \\ 0,030 & 0,31 & 0,31 & 0,31 & 0,30 \end{pmatrix}$$

Nous pouvons dès lors faire l'inventaire des propriétés intrinsèques de cette toute nouvelle matrice : dans un premier temps, $G$ est clairement stochastique, en tant qu'elle est combinaison convexe de deux matrices stochastiques. Nous perdons cependant du terrain par rapport aux matrices creuses $H$ et $S$, puisque $G$ est complèment dense.

Nonobstant, tout n'est pas perdu, puisque $G$ peut s'exprimer exclusivement en fonction de $H$ et de quelques vecteurs.

$$\text{En effet, } G = \alpha S + (1-\alpha)\frac{1}{N} e \cdot {}^t e = \alpha(H + \frac{1}{N} \mathrm{a} \cdot {}^t e) + (1-\alpha)\frac{1}{N} e \cdot {}^t e$$

$$\text{D'où } G = \alpha H + (\alpha \mathrm{a} + (1-\alpha)e)\frac{1}{N} {}^t e.$$

Ceci amènera des avantages non négligeables lorsqu'il faudra effectuer des calculs sur $G$.

$G$ possède de plus les propriétés essentielles à la convergence de notre vecteur invariant, parmi lesquelles la stochasticité et la stricte positivité (ceci sera plus amplement abordé dans la partie suivante). Ce sont ces derniers éléments, obtenus grâce aux périples de notre indispensable internaute aléatoire, qui vont assurer l'existence et l'unicité de ce non moins indispensable vecteur PageRank, et donc permettre la classification des pages web.

## 4. Vecteur invariant

Nous arrivons à présent à l'un des points les plus décisifs de la construction de Google. Certes la matrice G a été formée, de manière finalement plutôt simple, mais il est maintenant grand temps de justifier les ajustements réalisés à partir de la matrice $H$. Pourquoi une matrice stochastique ? Pourquoi des coefficients strictement positifs ? N'importe quel néophyte est en droit en se poser la question. Nous allons maintenant montrer que grâce à ces ajustements, il existe bien un et un seul vecteur invariant par G, ce qui permettra d'espérer un grand avenir pour le PageRank.

### 4.1. **Notations et définitions utilisées.**

Pour tout vecteur $V = \begin{pmatrix} v_1 & . & . & . & v_N \end{pmatrix} \in \mathcal{M}_{N,1}(\mathbb{R})$,
on note $|V| = \begin{pmatrix} |v_1| & . & . & . & |v_N| \end{pmatrix}$, et $||V||_1 = \sum_{i=1}^{N} |v_i|$.

Une matrice sera dite positive si tous ses coefficients sont des réels positifs ou nuls, strictement positive s'ils le sont strictement.

Un vecteur sera de probabilité si la somme de ses coefficients vaut 1.

Une matrice Q sera dite stochastique si elle est positive et si la somme de ses coefficients en ligne vaut 1 (i.e. si elle est composée de vecteurs-lignes de probabilité).

### 4.2. **Existence et unicité d'un vecteur de probabilité invariant.**

Notre but dans cette section sera de montrer l'existence et l'unicité d'un vecteur de probabilité invariant pour une matrice stochastique strictement positive.

Soit $Q \in \mathcal{M}_N(\mathbb{R})$ une matrice stochastique.
Soit $U$ l'élément de $\mathcal{M}_{1,N}(\mathbb{R})$ dont toutes les composantes valent 1.

Calculons $U \cdot {}^tQ = (n_j)$ en posant ${}^tQ = (q_{ij})$.

$n_i = \sum_{k=1}^{N}(1 \times q_{kj}) = \sum_{k=1}^{N} q_{kj} = 1$ car ${}^tQ$ est stochastique en colonne.

D'où, $\forall i, n_i = 1$, soit $U \cdot {}^tQ = U$.

$U \cdot {}^tQ = U$ donc $Q \cdot {}^tU = {}^tU$. Donc 1 est valeur propre de $Q$.
Montrons alors que $Q$ et ${}^tQ$ ont les mêmes valeurs propres, et considérons pour cela le polynôme caractéristique de $Q$, c'est à dire $\det(Q - X \cdot I_N)$.
On a $\det(Q - X \cdot I_N) = \det({}^t(Q - X \cdot I_N)) = \det({}^tQ - X \cdot I_N)$ donc $Q$ et ${}^tQ$ ont le même polynôme caractéristique, c'est à dire les mêmes valeurs propres. Or 1 est valeur propre de $Q$

Donc 1 est valeur propre de ${}^tQ$.

Soit $\lambda$=1, valeur propre de ${}^tQ$ et $V$ un vecteur propre de ${}^tQ$ associé à la valeur propre $\lambda$.
Prouvons que le vecteur ${}^tQ \cdot |V| - |V|$ est positif.
On a ${}^tQ \cdot V = V$, d'où $\forall i \in [\![1, n]\!], \sum_{j=1}^{N} {}^tQ(i,j)V(j) = V(i)$.

i.e. $|V(i)| = \left|\sum_{j=1}^{N} {}^tQ(i,j)V(j)\right| \leq \sum_{j=1}^{N} {}^tQ(i,j)|V(j)|, \qquad ({}^tQ(i,j) \geq 0)$

i.e. $\sum_{j=1}^{N} {}^tQ(i,j)|V(j)| - |V(i)| \geq 0$, $\forall i$.

D'où ${}^tQ|V| - |V|$ est positif.

Montrons alors que $|V|$ est invariant par ${}^tQ$.

On sait que la i-ième composante de ${}^tQ|V| - |V|$ est $(\sum_{j=1}^{N} Q(i,j)|V(j)|) - |V(i)|$.
Sommons ces composantes pour $i$ variant de 1 à $N$ : on obtient
$\sum_{i=1}^{N}((\sum_{j=1}^{N} {}^tQ(i,j)|V(j)|) - |V(i)|) = \sum_{i=1}^{N}\sum_{j=1}^{N} {}^tQ(i,j)|V(j)| - \sum_{i=1}^{N}|V(i)|$

$$= \sum_{j=1}^{N}(\sum_{i=1}^{N} {}^tQ(i,j))|V(j)| - \sum_{i=1}^{N}|V(i)| \qquad \left(\sum_{i=1}^{N} {}^tQ(i,j) = 1\right)$$
$$= \sum_{j=1}^{N}|V(j)| - \sum_{i=1}^{N}|V(i)| = 0$$

Or, $\forall\ i$, $(\sum_{j=1}^{N} {}^tQ(i,j)|V(j)|) - |V(i)| \geq 0$, donc

$$\left(\sum_{j=1}^{N} {}^tQ(i,j)|V(j)|\right) - |V(i)| \ = \ 0, \text{ i.e. } \left(\sum_{j=1}^{N} {}^tQ(i,j)|V(j)|\right) = |V(i)|$$

$\forall i$.

D'où ${}^tQ|V| = |V|$, i.e. $|V|$ est invariant par ${}^tQ$.

On en déduit que ${}^t|V|Q = {}^t|V|$.

Nous pouvons alors prouver l'existence d'au moins un vecteur de probabilité invariant par $Q$ pour la multiplication d'un vecteur-ligne par la matrice $Q$.
De fait, on sait qu'il existe un vecteur propre de ${}^tQ$ associé à 1. En prenant la valeur absolue de ce dernier, on obtient un nouveau vecteur, lui aussi invariant par ${}^tQ$.
Il suffit alors de diviser chacune de ses composantes par la somme de ses composantes pour en faire un vecteur de probabilité invariant.

On a donc bien prouvé l'existence d'un vecteur-ligne de probabilité invariant pour la multiplication à droite par $Q$.

Considérons alors une matrice $Q$ stochastique et strictement positive. Il existe donc au moins un vecteur de probabilité invariant par $Q$ noté

$$V_\infty = \big((v_\infty)_1 \quad . \quad . \quad . \quad (v_\infty)_N\big).$$

Nous allons maintenant prouver l'unicité de ce vecteur.

Montrons dans un premier temps que si $V$ est un vecteur positif invariant par $Q$, alors soit $V$ est nul, soit il est strictement positif.
Supposons au contraire qu'il existe $j_0$ tel que $v(j_0) = 0$.

Alors comme $V$ est invariant par $Q$, on peut écrire
$v(j_0) = \sum_{k=1}^{N} v(k)Q(k, j_0) = 0$. Or $Q(k, j_0) > 0 \ \forall k$ et $v(k) \geq 0 \ \forall k$.
Donc $v(k) = 0 \ \forall k$, soit $V = 0$.

Donc soit $V = 0$, soit $j_0$ n'existe pas et $V$ est strictement positif.

De plus, en supposant que $V_\infty$ n'est pas strictement positif, alors il est nul. Or la somme de de ses coefficients vaut 1. D'où une contradition. N'étant donc pas nul, on en déduit que $V_\infty$ est strictement positif.

On considère désormais un autre vecteur de probabilité noté $W_\infty$ invariant par Q. Puis on définit
$\alpha = \min \left\{ \frac{(w_\infty)_i}{(v_\infty)_i}, 1 \leq i \leq N \right\} = \frac{(w_\infty)_{i_0}}{(v_\infty)_{i_0}}$,
où $i_0$ est l'indice pour lequel ce minimum est atteint, et $V = W_\infty - \alpha V_\infty$.

Montrons que $V$ est invariant par $Q$.
$V \cdot Q = (W_\infty - \alpha V_\infty) \cdot Q = W_\infty \cdot Q - \alpha V_\infty \cdot Q = W_\infty - \alpha V_\infty = V$
par invariance de $V_\infty$ et $W_\infty$ par $Q$.

Donc $V$ est bien invariant par $Q$.

Nous pouvons alors prouver que $V$ est positif, sans toutefois l'être strictement.
En effet, $\forall j$, $V_j = (w_\infty)_j - \alpha(v_\infty)_j = (w_\infty)_j - \frac{(w_\infty)_{i_0}}{(v_\infty)_{i_0}}(v_\infty)_j$
$= (v_\infty)_j \cdot (\frac{(w_\infty)_j}{(v_\infty)_j} - \frac{(w_\infty)_{i_0}}{(v_\infty)_{i_0}})$. Par définition de $i_0$, on obtient $V_j \geq 0$,
donc $V$ est positif.
Cependant, la formule précédemment obtenue montre que $V(i_0) = 0$.

Donc $V$ n'est pas strictement positif.

$V$ est invariant par $Q$, positif mais pas strictement positif,
donc V est nul, soit $W_\infty - \alpha V_\infty = 0$.

D'où $W_\infty = \alpha V_\infty$

Qui plus est, $V_\infty$ et $W_\infty$ sont des vecteurs de probabilité, donc la somme de leurs coefficients est égale à 1. Il faut donc nécessairement $\alpha = 1$ pour vérifier l'égalité.

$$\text{Donc } V_\infty = W_\infty$$

### 4.3. **Convergence de la suite $(Q^n)_{n\geq 1}$.**

Dans cette partie, nous nous attacherons à montrer que les puissances successives de la matrice $G$ convergent vers une matrice dont tous les vecteurs sont égaux au vecteur invariant que nous cherchons.
Nous allons pour cela considérer une matrice $Q$ stochastique strictement positive et son vecteur de probabilité invariant $V_\infty$.
Dans un premier temps, nous allons localiser les valeurs propres de $Q$.
Pour cela, commençons par vérifier que pour tout vecteur $V \in \mathcal{M}_{1,N}(\mathbb{R})$ on a : $||V \cdot Q||_1 \leq ||V||_1$.

$$\text{En effet, } ||V \cdot Q||_1 = \sum_{j=1}^{N} \left| \sum_{k=1}^{N} V(k)Q(k,j)) \right| \leq \sum_{k=1}^{N}\sum_{j=1}^{N} |V(k)|Q(k,j)$$

$$\text{Or } \sum_{j=1}^{N} Q(i,j) = 1. \text{ d'où } ||V \cdot Q||_1 \leq \sum_{j=1}^{N} |V(j)| = ||V||_1$$

$$\text{On a bien } ||V \cdot Q||_1 \leq ||V||_1.$$

Si de plus $V$ est positif, alors

$$||V \cdot Q||_1 = \sum_{j=1}^{N} \left| \sum_{k=1}^{N} V(k)Q(k,j)) \right| = \sum_{k=1}^{N}\sum_{j=1}^{N} |V(k)|Q(k,j) = ||V||_1.$$

$$\text{On a alors } ||V \cdot Q||_1 = ||V||_1.$$

Nous pouvons alors en déduire que pour toute valeur propre réelle $\lambda$ de $Q$, on a $|\lambda| \leq 1$.
En effet, soit $\lambda$ une valeur propre de $Q$. $\lambda$ est également valeur propre de ${}^tQ$, donc $\exists\ V \in \mathcal{M}_{1,N}(\mathbb{R})$, ${}^tQ \cdot {}^tV = \lambda {}^tV$, i.e. $V \cdot Q = \lambda V$.
$||V \cdot Q||_1 = ||\lambda V||_1 = |\lambda| \times ||V||_1$, d'où $|\lambda| \times ||V||_1 \leq ||V||_1$.

$$\text{D'où finalement } |\lambda| \leq 1.$$

Considérons désormais une valeur propre réelle $\lambda$ telle que $|\lambda| = 1$, et $V$ un vecteur propre de $Q$ associé à $\lambda$ tel $||V||_1 = 1$. On sait grâce aux résultats précédents que $|V| = V_\infty$.
Etablissons l'égalité suivante :

$$\left|\sum_{k=1}^{N} Q(1,k)v_k\right| = \sum_{k=1}^{N} Q(1,k)|v_k|$$

On a $\lambda = \pm 1$, donc $Q \cdot V = \pm V$, i.e. $Q \cdot |V| = |V|$, d'où
$\left|\sum_{k=1}^{N} Q(1,k)v_k\right| = |V_1| = \sum_{k=1}^{N} Q(1,k)|v_k|$. D'où le résultat.

Or on montre facilement, à l'aide de manipulations sur les valeurs absolues, que si cette égalité est vérifiée, alors $V = |V|$ ou $V = -|V|$.

On a $Q \cdot |V| = |V|$, donc $Q \cdot \varepsilon|V| = \varepsilon|V|$, avec $V = \varepsilon|V|$, $\varepsilon = \pm 1$.
Donc $Q \cdot V = V$.

On obtient donc finalement $\lambda = 1$.

On se placera ici dans le cas où $Q$ est diagonalisable. Il existe alors une matrice diagonale $D$ et une matrice inversible $S$ telles que $\forall n \geq 1$, $Q^n = S \cdot D^n \cdot S^{-1}$.
(En effet, $Q$ étant diagonalisable, ce résultat est vérifié pour $n = 1$.
$Q^n$ s'écrit alors $Q^n = (S \cdot D \cdot S^{-1}) \cdot (S \cdot D \cdot S^{-1}) \cdot ... \cdot (S \cdot D \cdot S^{-1})$,
d'où, par associativité du produit matriciel, le résultat).

Nous pouvons alors montrer que parmi les $N$ coefficients diagonaux de $D$, tous sont de valeur absolue strictement inférieure à 1, sauf un qui y est égal.
On sait que pour toute valeur propre de $Q$ on a $|\lambda| \leq 1$, et de plus, 1 est valeur propre simple.

On en déduit qu'il n'y a effectivement qu'un seul 1 sur la diagonale, les autres coefficients étant de module strictement inférieur à 1.

Ce dernier résultat prouve que $(Q^n)_{n\geq 1}$ converge vers une matrice $Q_\infty$.
Notons $D = (d_i)$, et $S = (s_{ij})$. $D^n = (d_i^n)$, d'où $S \cdot D^n = (s_{ij} \times d_i^n)$.
Notons $i_0$ l'entier tel que $d_{i_0} = 1$. Alors $\forall i \neq d_{i_0}$, on a $s_{ij} \times d_i^n \to 0$, et $s_{i_0 j} \times d_{i_0}^n \to 1$. Donc $S \cdot D^n$ converge, et de là, $S \cdot D^n \cdot S^{-1}$ aussi.

$$\text{Donc } \exists\ Q_\infty,\ \lim_{v \to +\infty} Q^n = Q_\infty.$$

Montrons par ailleurs que le produit de deux matrices stochastiques est une matrice stochastique.

Soient $A = (a_{ij})$, $B = (b_{ij})$ et $C = A \cdot B = (c_{ij})$. $c_{ij} = \sum_{k=1}^{N} a_{ik} b_{kj}$.

Donc $\sum_{j=1}^{N} c_{ij} = \sum_{j=1}^{N} \sum_{k=1}^{N} a_{ik} b_{kj} = \sum_{k=1}^{N} a_{ik} \left( \sum_{j=1}^{N} b_{kj} \right) = \sum_{k=1}^{N} a_{ik} = 1.$

Ainsi, comme $Q$ est stochastique, ses puissances le sont également, et en particulier $Q_\infty$ l'est.

De plus, comme $Q^n \cdot Q = Q^{n+1}$, on obtient, par passage à la limite quand $n \to +\infty$, $Q_\infty \cdot Q = Q_\infty$, donc $Q_\infty$ est invariante par $Q$.

Donc les vecteurs-lignes de $Q_\infty$ sont invariants par $Q$. Ce sont des vecteurs de probabilité. L'unicité du vecteur invariant indique alors que

Chacun des vecteurs de $Q_\infty$ est égal à $V_\infty$.

Tout semble donc se passer à merveille dans les premières vérifications des propriétés indispensables du vecteur PageRank. En effet, les caractéristiques de la matrice $G$, notamment sa stochasticité et la positivité de ses coefficients, permettent au vecteur invariant non seulement d'exister, mais également d'être unique. C'est l'un des points cruciaux pour le but que Google s'est fixé, car imaginez un instant que le vecteur PageRank ne soit pas unique... Quel classement de pages choisir alors ? Qui plus est, outre ces caractéristiques salutaires, une autre propriété de la matrice Google est apparue : celle de converger vers une matrice présentant le vecteur PageRank lorsqu'on la multiplie par elle-même. Nous verrons en partie 6 que cette propriété n'est pas celle utilisée pour déterminer le PageRank ; néanmoins, elle reste l'une des particularités qui font toute la magie du PageRank.

## 5. Dépendance par rapport au paramètre $\alpha$

La matrice Google se présente donc sous une forme très simple : $G = \alpha S + (1-\alpha)E$. Cette expression semble de plus être parfaitement adaptée au problème du PageRank car, comme nous l'avons vu dans le chapitre précédent, elle garantit l'existence et l'unicité du vecteur invariant qui ordonne les pages du Web. Cependant, si les paramètres $S$ et $E$ ne posent aucune difficulté, car sont directement issus de la structure du graphe du Web, ce n'est pas la même chose pour $\alpha$.

Google utilise la valeur $\alpha = 0.85$ . Il faut ainsi se demander en quoi cette valeur apparaît-elle comme la plus raisonnable par rapport à la détermination du PageRank. Comment dépendent donc les différents éléments qui entrent en jeu par rapport à ce mystérieux paramètre $\alpha$ ?

### 5.1. **Que représente $\alpha$ ?**

Comme cela est expliqué dans les précedents chapitres, $\alpha$ représente en fait une probabilité. Effectivement, $S$ est la matrice représentant le graphe du Web, et $E$ est la matrice de téléportation. Un surfer aléatoire aurait ainsi à chaque étape deux possibilités : la première serait d'emprunter un lien menant à une nouvelle page, l'autre de se rendre directement à une page quelconque, sans utiliser le graphe du Web. Le paramètre $\alpha$ symbolise donc la probabilité que ce surfer aléatoire a d'utiliser les liens entre les pages.

Il semble alors logique de considérer les valeurs de $\alpha$ proches de 1. En effet, le but du PageRank est tout d'abord de rendre compte de l'importance d'une page, c'est-à-dire de son importance dans le graphe du Web, par rapport aux liens qui la relient aux autres pages. Ainsi, il semble logique de privilégier la matrice $S$, qui représente les liens entre pages, plutôt que cette matrice $E$ qui envisage toutes les pages avec équiprobabilité.

$\alpha = 1$ est toutefois à éliminer, car nous avons vu que la matrice $E$ est nécessaire pour l'existence du vecteur PageRank. Cependant, pourquoi Page et Brin, les créateurs de Google, n'ont-ils pas choisi $\alpha = 0.99$ plutôt que $\alpha = 0.85$ ?

## 5.2. **Rapidité de la convergence.**

L'une des questions les plus importantes concernant le fonctionnement de la méthode de classement des pages est celle de la vitesse de convergence. En effet, il s'agit de traiter des milliards de pages et l'idéal est donc d'obtenir un vecteur invariant proche de la limite en un nombre faible d'opérations.

Nous verrons en partie 6 que la méthode de calcul du PageRank utilisée par Google est celle dite de la puissance. Cette méthode est une méthode itérative qui, quand le nombre d'opérations tend vers l'infini, converge vers le vecteur invariant recherché (ce que nous montrerons au chapitre consacré à cette méthode), ce qui est sans nul doute le point le plus appréciable dans le but de la méthode. La vitesse de convergence correspond dans ce cas au nombre d'itérations nécessaires pour avoir une assez bonne précision.

En fait, $\alpha$ est en grande partie responsable de cette vitesse de convergence. En effet, cette vitesse de convergence dépend essentiellement de la vitesse à laquelle $|\lambda_2|^k$ tend vers 0 ($\lambda_2$ étant la deuxième valeur propre de $G$ par ordre de modules décroissants ; cf chapitre 6 : la méthode de la puissance, pour une démonstration de ce phénomène). Plus $\lambda_2$ sera donc proche de 0, plus le vecteur invariant pourra être calculé rapidement. Mais $\lambda_2$ dépend en réalité de $\alpha$, car si l'on appelle $\{1, \mu_2, ..., \mu_n\}$ l'ensemble des valeurs propres de $S$, et $\{1, \lambda_2, ..., \lambda_n\}$ celui de $G$, alors $\lambda_k = \alpha\mu_k$ pour tout $k$ entre 2 et $n$. Montrons ce résultat.

Pour cela, posons $\hat{e} = \dfrac{e}{\sqrt{N}}$, et complétons $(\hat{e})$ en une

base orthonormale de $\mathbb{R}^{\mathbb{N}}$ pour la structure euclidienne canonique.

$\exists\, U_1 \in \mathcal{M}_{N,N-1}(\mathbb{R}), (\hat{e}, U_1) \in \mathcal{O}_N(\mathbb{R})$. Posons $U = (\hat{e}\ \ U_1)$.

$U$ est donc la matrice de passage de la base canonique à la base que nous venons de construire.
Etudions ${}^tU \cdot {}^tS \cdot U = S_1$. On a $S \cdot \hat{e} = \dfrac{1}{\sqrt{N}} S \cdot e = \dfrac{e}{\sqrt{N}} = \hat{e}$

car S est stochastique en ligne. D'où $S_1 = {}^tU \cdot {}^tS \cdot U = \begin{pmatrix} 1 & 0 & . & . & . & 0 \\ w & & & T & & \end{pmatrix}$.

car ${}^tS_1$ représente $S$ dans la base orthonormale construite.

Or $\mathcal{X}_{{}^tS} = \det({}^tS - X \cdot I_N) = \det(U \cdot S_1 \cdot {}^tU - X \cdot I_N)$

$= \det(U \cdot (S_1 - X \cdot I_N) \cdot {}^tU) = \det(S_1 - X \cdot I_N) = \mathcal{X}_{S_1}$.

D'où $\mathcal{X}_{{}^tS} = \begin{vmatrix} 1-X & 0 \\ W & T - X \cdot I_{N-1} \end{vmatrix} = (1-X)\det(T - X \cdot I_{N-1})$

Soit $\mathcal{X}_{{}^tS} = (1-X)\mathcal{X}_T$.
Comme 1 est valeur propre de $S$, $\mu_2, ..., \mu_N$ sont les racines de $\mathcal{X}_T$.

Considérons maintenant ${}^tU \cdot {}^tG \cdot U = {}^tU \cdot (\alpha {}^tS + (1-\alpha)\frac{1}{N} e \cdot {}^te) \cdot U$

$= \alpha {}^tU \cdot {}^tS \cdot U + \frac{1-\alpha}{N} {}^tU \cdot e \cdot {}^te \cdot U$.

Or ${}^teU = {}^te \cdot (\hat{e} \;\; U_1) = (\sqrt{N} \;\; 0 \;\; ... \;\; 0)$ car les vecteurs-colonnes de $U_1$ sont orthogonaux à $\hat{e}$, donc à $e$.

D'où ${}^tU \cdot e \cdot {}^te \cdot U = \begin{pmatrix} N & 0 & . & . & . & 0 \\ 0 & 0 & . & . & . & 0 \\ . & . & & & & . \\ . & . & & & & . \\ 0 & 0 & . & . & . & 0 \end{pmatrix}$

i.e. ${}^tU \cdot G \cdot U = \begin{pmatrix} \alpha & 0 & . & . & . & 0 \\ \alpha w & & & \alpha T & & \end{pmatrix} + (1-\alpha) \begin{pmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 0 & . & . & . & 0 \\ . & . & & & & . \\ . & . & & & & . \\ 0 & 0 & . & . & . & 0 \end{pmatrix}$

$= \begin{pmatrix} 1 & 0 & . & . & . & 0 \\ \alpha w & & & \alpha T & & \end{pmatrix}$.

D'où $\mathcal{X}_G = \mathcal{X}_{{}^tU \cdot G \cdot U} = (1-X)\det(\alpha T - X \cdot I_N)$.
Les racines de $\mathcal{X}_G$ sont donc 1 et les racines de l'équation

$\det(\alpha T - X \cdot I_N) = 0$, i.e. $\alpha^{N-1}\det(T - \frac{1}{\alpha} X \cdot I_N) = 0$.

Donc $\frac{1}{\alpha} X \in Sp(T)$, i.e. $\frac{1}{\alpha} X \in \{\mu_2, ..., \mu_N\}$.

Le spectre de $G$ est donc $\{1, \alpha\mu_1, ..., \alpha\mu_N\}$

On a donc bien $\lambda_k = \alpha\mu_k$, $\forall\ k \leq 2$.

Ainsi, $\lambda_2$ commandant la convergence, c'est $\alpha\mu_2$ qui la rend plus ou moins forte. Cependant, $\mu_2$ est la deuxième valeur propre de $S$ par ordre de module décroissant, son module est donc très proche de 1. Finalement, plus $\alpha^k$ tend vers 0 rapidement, plus le PageRank est proche de la réalité rapidement. Il semble donc falloir utiliser des valeurs faibles de $\alpha$ pour ne pas faire un trop grand nombre d'itérations pour déterminer le vecteur invariant.

Pour donner quelques exemples, il faut seulement 34 itérations pour obtenir le vecteur invariant avec une précision de $10^{-10}$, lorsque $\alpha = 0.5$. Dans les mêmes conditions, il en faut 142 pour $\alpha = 0.85$, et 23015 pour $\alpha = 0.999\,!$

Rappelons tout de même que pour que la matrice $G$ ait la signification qu'on veut lui donner, il faut choisir $\alpha$ assez proche de 1. Il faut donc trouver le meilleur compromis pour que la structure du Web ait un sens dans la manière dont elle intervient dans le classement des pages, mais également pour que la vitesse de convergence soit assez élevée.

### 5.3. **Variations du PageRank avec $\alpha$.**

Dans le cadre de l'utilisation d'un moteur de recherche tel que Google, le classement des pages se doit d'être le moins sensible possible à d'éventuelles variations des paramètres entrant en jeu dans son calcul. En effet, si le classement était modifié en permanence, il n'aurait plus aucun sens, et ne représenterait pas du tout la réalité. Dans cette partie, nous étudierons la dépendance du vecteur PageRank par rapport à $\alpha$. Pour cela, nous établirons une série de résultats concernant $\frac{d\pi(\alpha)}{d\alpha}$, dérivée du vecteur par rapport à $\alpha$.

Mais avant tout calcul concernant cette dérivée, il s'agit de montrer que $\pi(\alpha)$ est bien dérivable sur ]0,1[, sans quoi la suite n'aurait aucun sens. Pour cela, déterminons une expression de $\pi(\alpha)$ qui prouve la dérivabilité, en montrant que

$$\pi(\alpha) = \frac{1}{\sum_{i=1}^{n} D_i(\alpha)}(D_1(\alpha), D_2(\alpha), ..., D_n(\alpha)),$$

où $D_k(\alpha)$ est le k-ième mineur diagonal d'ordre $n-1$ de la matrice $A = I - G$.

*Preuve :* Posons $A = I - G$.

On sait que $A \cdot \tilde{A} = \tilde{A} \cdot A = \det(A) \cdot I_n$.
Or 1 est valeur propre de $G$.
Donc $\exists\ X$, $G \cdot X = X$, c'est à dire $(I - G) \cdot X = 0$, i.e. $A \cdot X = 0$.
Comme $X \neq 0$, on a $A \notin GL_n(\mathbb{K})$. Donc $\det(A) = 0$, c'est à dire

$$A \cdot \tilde{A} = \tilde{A} \cdot A = 0.$$

Or $Ker(A) = Ker(I - G)$, et 1 est valeur propre de $G$.
Et $\dim(Ker(I-G)) = 1$ car $\dim(E_1(G)) = 1$. Donc $\dim(Ker(A)) = 1$.
Le théorème du rang amène alors :

$$\mathrm{rg}(A) = n - 1.$$

Comme de plus $\tilde{A} \cdot A = 0$, alors $\forall\ X \in \mathcal{M}_{n,1}(\mathbb{K}), \tilde{A} \cdot (A \cdot X) = 0$. C'est à dire $Im(A) \subset Ker(\tilde{A})$, d'où $\dim(Ker(\tilde{A})) \geq \mathrm{rg}(A) = n - 1$ et $\tilde{A} \neq 0$, donc $\dim(Ker(\tilde{A})) = n - 1$. On en déduit que

$$\mathrm{rg}(\tilde{A}) = 1.$$

Notons $\tilde{A} = (\tilde{A}_1\ \tilde{A}_2\ ...\ \tilde{A}_n)$. $\tilde{A}_i$ étant un vecteur-colonne.
On a $A \cdot \tilde{A} = 0 \ \Leftrightarrow\ (A \cdot \tilde{A}_1\ A \cdot \tilde{A}_2\ ...\ A \cdot \tilde{A}_n) = 0 \ \Leftrightarrow \forall\ i, A \cdot \tilde{A}_i = 0$.
D'où $\forall i$, $\tilde{A}_i \in Ker(A)$. Or $\dim(Ker(A)) = 1$ et $e \in Ker(A)$
car $(I - G) \cdot e = e - G \cdot e = 0$, $G$ étant stochastique.
Donc les $\tilde{A}_i$ sont tous colinéaires à e, c'est à dire qu'en posant
$\tilde{A}_j = w_j \cdot e$, on obtient $\tilde{A} = e \cdot (w_1\ ...\ w_n)$. Or $\tilde{A}_{ii} = D_i$, le cofacteur ayant un signe + car le coefficient est sur la diagonale.

Comme on a $\tilde{A} = \begin{pmatrix} w_1 & . & . & . & w_n \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ w_1 & . & . & . & w_n \end{pmatrix}$, alors $w_i = \tilde{A}_{ii} = D_i$.

$$\text{Donc } (w_1\ ...\ w_n) = (D_1\ ...\ D_n).$$

De même, $\tilde{A} \cdot A = 0$, d'où, en notant $\tilde{A} = \begin{pmatrix} \tilde{A}'_1 \\ . \\ . \\ . \\ \tilde{A}'_n \end{pmatrix}$, avec $\tilde{A}'_i = w$,

où $w = \begin{pmatrix} w_1 & ... & w_n \end{pmatrix}$, on obtient :

${}^tA \cdot {}^t\tilde{A} = 0$, donc $\forall i$, ${}^tA \cdot {}^t\tilde{A}'_i = 0$. Or ${}^tA \cdot {}^t\pi = 0$ et $\text{rg}(A) = \text{rg}({}^tA) = n-1$.
Donc $\exists \lambda$, ${}^t\tilde{A}'_i = \lambda {}^t\pi$, i.e. ${}^tw = \lambda {}^t\pi$.

$$\text{D'où } w = \lambda\pi,$$

et $\lambda \neq 0$ car sinon, on aurait $w = 0$, d'où $\tilde{A} = 0$, ce qui est impossible car $\text{rg}(\tilde{A}) = 1$.
Donc $w = \lambda\pi$, c'est à dire $w \cdot e = \lambda\pi \cdot e = \lambda$ car $\pi \cdot e = 1$ ($\pi$ est une distribution de probabilité). On en déduit que $\dfrac{w}{w \cdot e} = \dfrac{w}{\lambda} = \pi$.
Finalement, on a bien

$$\pi(\alpha) = \frac{(D_1(\alpha), D_2(\alpha), ..., D_n(\alpha))}{\sum_{i=1}^{n} D_i(\alpha)}.$$

Ainsi, chaque composante de $\pi(\alpha)$ s'exprimant en fonction de somme et de produits d'une matrice connue (car un mineur n'est rien d'autre qu'un déterminant, c'est-à-dire une somme de produits de coefficients), $\pi(\alpha)$ est bien dérivable par rapport à $\alpha$ lorsque $\alpha$ varie dans ]0,1[. Nous sommes donc en droit d'étudier les variations de $\pi(\alpha)$ grâce à sa dérivée.
Commençons par déterminer une majoration de $|\frac{d\pi(\alpha)}{d\alpha}|$, et notons pour cette démonstration $\pi(\alpha) = (\pi_1(\alpha), \pi_2(\alpha), ..., \pi_n(\alpha))$.

On a $\pi \cdot e = 1$ car $\pi$ est une distribution de probabilité.
Donc $\dfrac{d\pi}{d\alpha} \cdot e + \pi \cdot \dfrac{de}{d\alpha} = \dfrac{d}{d\alpha}(1)$, soit

$$\frac{d\pi}{d\alpha} \cdot e = 0.$$

On a de plus, par invariance de $\pi$, $\pi(\alpha) = \pi(\alpha)\cdot(\alpha S+(1-\alpha)e\cdot {}^t v)$, d'où, en dérivant : $\dfrac{d\pi(\alpha)}{d\alpha} = \dfrac{d\pi(\alpha)}{d\alpha}\cdot(\alpha S + (1-\alpha)e\cdot {}^t v) + \pi(\alpha)\cdot(S - e\cdot {}^t v)$

$= \alpha\dfrac{d\pi(\alpha)}{d\alpha}S + (1-\alpha)\dfrac{d\pi(\alpha)}{d\alpha}e\cdot {}^t v + \pi(\alpha)\cdot(S - e\cdot {}^t v)$

$= \alpha\dfrac{d\pi(\alpha)}{d\alpha}S + \pi(\alpha)\cdot(S - e\cdot {}^t v)$ car $\dfrac{d\pi(\alpha)}{d\alpha}e = 0.$

$$\text{D'où } \frac{d\pi(\alpha)}{d\alpha}(I - \alpha S) = \pi(\alpha)\cdot(S - e\cdot {}^t v).$$

Montrons alors que $I - \alpha S$ est inversible.

On a : $||\alpha S(\alpha)||_\infty^k \leq \alpha^k ||S(\alpha)||_\infty^k = (\alpha||S(\alpha)||_\infty)^k$. Ici, $\alpha||S(\alpha)||_\infty < 1$. En effet, la somme des coefficients de $S$ en ligne vaut 1, car S est stochastique en ligne, d'où $\alpha||S(\alpha)||_\infty = \alpha < 1$.
La série $\displaystyle\sum_{k\geq 0} ||\alpha S(\alpha)||_\infty^k$ converge donc. Comme l'absolue convergence entraîne la convergence, la série $\displaystyle\sum_{k=0}^{N}(\alpha S(\alpha))^k$ converge également.

Or on a $\displaystyle\sum_{k=0}^{N}(\alpha S(\alpha))^k \times (I - \alpha S) = I - \alpha^{N+1}S(\alpha)^{N+1} \underset{N\to+\infty}{\longrightarrow} I$ car la somme est téléscopique.

D'où $\displaystyle\sum_{k=0}^{+\infty}(\alpha S(\alpha))^k \times (I - \alpha S) = I.$
La matrice $I - \alpha S$ est donc inversible.

$$\text{On a alors } \frac{d\pi(\alpha)}{d\alpha} = \pi(\alpha)\cdot(S - e\cdot {}^t v)\cdot(I - \alpha S)^{-1}.$$

Soit maintenant $x \in e^\perp$ pour le produit scalaire adéquat. ${}^t x\cdot e = 0$, d'où, $\forall y \in \mathcal{M}_{n,1}(\mathbb{R})$, $|{}^t x\cdot y| = |{}^t x\cdot(y - \alpha e)| = \left|\displaystyle\sum_{i=1}^{n} x_i y_i'\right| \leq \sum_{i=1}^{n}|x_i||y_i'|$

en posant ${}^t x = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$ et $y - \alpha e = \begin{pmatrix} y_1' \\ . \\ . \\ . \\ y_n' \end{pmatrix}$.

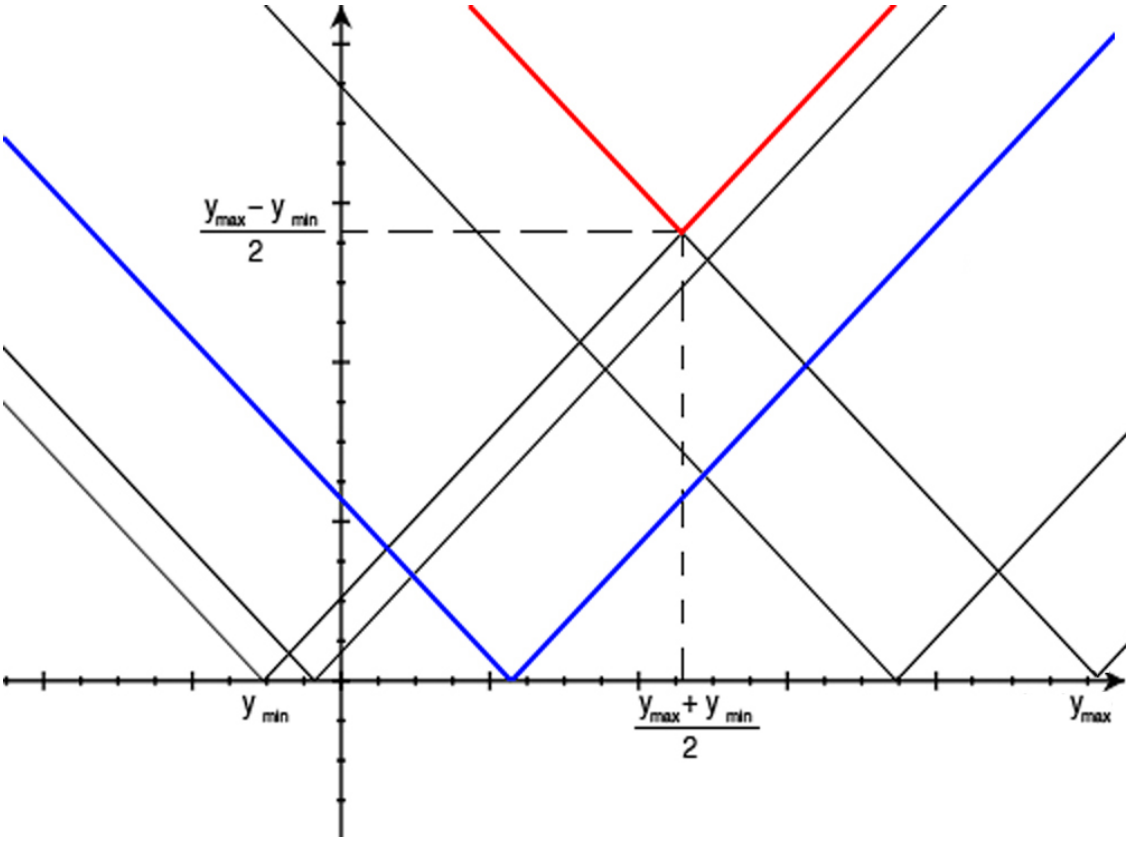
Comme $||y - \alpha e||_\infty = \max(|y'_1|, ..., |y'_n|)$, on a

$$|^t x \cdot (y - \alpha e)| \leq ||y - \alpha e||_\infty \left( \sum_{i=1}^{n} |x_i| \right) = ||x||_1 ||y - \alpha e||_\infty.$$

Etudions maintenant $||y - \alpha e||_\infty$.

$y - \alpha e = \begin{pmatrix} y_1 \\ . \\ . \\ . \\ y_n \end{pmatrix} - \alpha \begin{pmatrix} 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}$, on a $||y - \alpha e||_\infty = \max(|y_1 - \alpha|, ..., |y_n - \alpha|)$.

Or, le graphique suivant le montre (la courbe rouge représente $\max(|y_1 - \alpha|, ..., |y_n - \alpha|$ ; la bleue représente une fonction $|y_k - \alpha|$),



ce maximum est atteint pour $\alpha = \frac{y_{min} + y_{max}}{2}$, et vaut

$$\frac{y_{max} - y_{min}}{2}.$$

$$\text{D'où } |{}^t x \cdot y| \leq ||x||_1 \frac{y_{max} - y_{min}}{2}.$$

On a de plus : $\frac{d\pi_j(\alpha)}{d\alpha} = \frac{d\pi(\alpha)}{d\alpha} e_j$, où $e_j$ est le j-ième vecteur colonne de $I_n$.

$$\text{D'où } \frac{d\pi_j(\alpha)}{d\alpha} = \pi(\alpha) \cdot (S - e \cdot {}^t v) \cdot (I - \alpha S)^{-1} e_j$$

Montrons alors que $\pi(\alpha) \cdot (S - e \cdot {}^t v) \cdot e = 0$, en vue d'appliquer l'inégalité précédemment établie à $y = (I - \alpha S)^{-1} e_j$. Posons $S = (s_{ij})$ et $x = {}^t(\pi(\alpha) \cdot (S - e \cdot {}^t v))$.
$S - e \cdot {}^t v = (s_{ij} - \frac{1}{n})$, donc

$$(S - e \cdot {}^t v) \cdot e = \begin{pmatrix} s_{11} - \frac{1}{n} & . & . & . & s_{1n} - \frac{1}{n} \\ . & & & & . \\ . & & & & . \\ . & & & & . \\ s_{n1} - \frac{1}{n} & . & . & . & s_{nn} - \frac{1}{n} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ . \\ . \\ . \\ 1 \end{pmatrix} = \begin{pmatrix} s_{11} + ... + s_{1n} - 1 \\ . \\ . \\ . \\ s_{n1} + ... + s_{nn} - 1 \end{pmatrix}$$

D'où $\pi(\alpha) \cdot (S - e \cdot {}^t v) \cdot e = \sum_{k=1}^{n} \pi_k(\alpha)(s_{k1} + ... + s_{kn} - 1) = 0$ car S est stochastique.

En appliquant l'inégalité avec ${}^t x = \pi(\alpha) \cdot (S - e \cdot {}^t v)$ et $y = (I - \alpha S)^{-1} \cdot e_j$ (on vient de monter que $x \in e^{\perp}$),

on a alors : $\left| \frac{d\pi_j(\alpha)}{d\alpha} \right| \leq ||\pi(\alpha)(S - e \cdot {}^t v)||_1 \frac{y_{max} - y_{min}}{2}$.

Or $||\pi(\alpha)(S - e \cdot {}^t v)||_1 \leq 2$. En effet $||\pi(\alpha)(S - e \cdot {}^t v)||_1 = ||\pi_i(\alpha)(s_{ij} - \frac{1}{n})||_1$

$= \sum_{j=1}^{n} \left| \sum_{i=1}^{n} \pi_i(\alpha)(s_{ij} - \frac{1}{n}) \right| = \sum_{j=1}^{n} \left| \left( \sum_{i=1}^{n} \pi_i(\alpha) s_{ij} \right) - \frac{1}{n} \right|$ car $\pi$ est une distribution de probabiblité.

Ceci est inférieur à $\left(\sum_{j=1}^{n}\sum_{i=1}^{n}\pi_i(\alpha)s_{ij}\right)+1=2$
par stochasticité de S : en factorisant par la somme des $\pi_i(\alpha)$, on trouve le produit de la somme des $s_{ij}$ et de la somme des $\pi_i(\alpha)$.

$$\text{D'où } \left|\frac{d\pi_j(\alpha)}{d\alpha}\right| \leq y_{max} - y_{min}$$

Or on a vu que $(I-\alpha S)^{-1} = \sum_{k=0}^{+\infty}\alpha^k S^k$
et $(I-\alpha S)\cdot e = e - \alpha S\cdot e = (1-\alpha)e$
car S est stochastique donc $S\cdot e = e$.
Donc $(I-\alpha S)^{-1}\cdot e = \frac{1}{(1-\alpha)}e$.

$$\text{Donc } y_{min} = ((I-\alpha S)^{-1}e_j)_{min} \geq 0.$$

Quant à $y_{max}$, on a :

Soit $(I-\alpha S)^{-1} = (\beta_{ij})$. On a $(I-\alpha S)^{-1}\cdot e = \begin{pmatrix}\beta_{11}+...+\beta_{1n}\\ .\\ .\\ .\\ \beta_{n1}+...+\beta_{nn}\end{pmatrix}$

et les $\beta_{ij}$ sont tous positifs, donc $\beta_{ij} \leq \beta_{i1}+...+\beta_{in} \leq ||(I-\alpha S)^{-1}||_\infty$
donc $\max_{i,j}\beta_{ij} \leq ||(I-\alpha S)^{-1}||_\infty$, c'est à dire :

$$y_{max} \leq ||(I-\alpha S)^{-1}e||_\infty = ||\frac{1}{1-\alpha}e||_\infty$$

$$\text{D'où } y_{max} \leq \frac{1}{1-\alpha}.$$

On a donc : $\left|\frac{d\pi_j(\alpha)}{d\alpha}\right| \leq y_{max} - y_{min}$, $y_{min} \geq 0$ et $y_{max} \leq \frac{1}{1-\alpha}$.

$$\text{Finalement, } \left|\frac{d\pi_j(\alpha)}{d\alpha}\right| \leq \frac{1}{1-\alpha}.$$

Cette majoration nous donne donc d'importantes informations concernant les variations du PageRank avec $\alpha$, notamment pour les faibles valeurs de $\alpha$. En effet, plus $\alpha$ est proche de 0, plus le majorant est faible. Ainsi, lorsque $\alpha$ est petit, $\frac{d\pi(\alpha)}{d\alpha}$ a une faible valeur absolue, ce qui signifie que $\pi(\alpha)$ ne varie que très peu lorsque $\alpha$ varie. Les petites valeurs de $\alpha$ semblent donc convenir pour que le PageRank soit assez insensible aux variations de $\alpha$.

Cependant, il est vrai que pour les grandes valeurs de $\alpha$, cette majoration est inutile. En effet, $\frac{1}{1-\alpha} \to +\infty$ quand $\alpha$ tend vers 1. Le majorant est donc trop grand pour que l'inégalité donne un renseignement qui ait un sens concernant l'ordre de grandeur de $|\frac{d\pi(\alpha)}{d\alpha}|$.

Déterminons alors une expression de $\frac{d\pi(\alpha)}{d\alpha}$ qui nous permettra d'étudier les limites lorsque $\alpha \to 0$, ou lorsque $\alpha \to 1$ (les résultats concernant les groupes multiplicatifs de matrices et les inverses associés se trouvent dans l'appendice).

On a $\pi(\alpha) = \pi(\alpha) \cdot G = \pi(\alpha) \cdot (\alpha S + (1-\alpha) \cdot e^t v)$ par invariance du PageRank par $G$. D'où : $\pi(\alpha) \cdot (I - \alpha S - (1-\alpha) \cdot e^t v) = 0$.
On a vu que $I - \alpha S$ était inversible.
On peut donc multiplier les deux membres de l'égalité par $(I - \alpha S)^{-1}$.
On obtient
$\pi(\alpha) \cdot ((I - \alpha S) - (1-\alpha) \cdot e \cdot {}^t v) \cdot (I - \alpha S)^{-1} = 0$
Soit $\pi(\alpha) \cdot (I - (1-\alpha) \cdot e \cdot {}^t v \cdot (I - \alpha S)^{-1}) = 0$
Donc $\pi(\alpha) = (1-\alpha) \cdot {}^t v \cdot (I - \alpha S)^{-1}$ car $\pi(\alpha) \cdot e = 1$.

D'où $\frac{d\pi(\alpha)}{d\alpha} = -{}^t v \cdot (I - \alpha S)^{-1} - (1-\alpha) \cdot {}^t v \cdot (-(I - \alpha S)^{-1} \cdot S \cdot (I - \alpha S)^{-1})$

en dérivant le produit.

ie : $\frac{d\pi(\alpha)}{d\alpha} = (1-\alpha)^t v \cdot (I - \alpha S)^{-1} \cdot S \cdot (I - \alpha S)^{-1} - {}^t v \cdot (I - \alpha S)^{-1}$

$= -{}^t v (I - \alpha S)^{-1} \cdot (I - (1-\alpha) S \cdot (I - \alpha S)^{-1})$

$= -{}^t v (I - \alpha S)^{-1} \cdot ((I - \alpha S) - (1-\alpha) S) \cdot (I - \alpha S)^{-1}$

$= -{}^t v (I - \alpha S)^{-1} \cdot (I - S) \cdot (I - \alpha S)^{-1}$.

Or $(I - \alpha S)^{-1}$ et $(I - S)$ commutent.

En effet, $(I-\alpha S)^{-1} = \sum_{k=0}^{+\infty} \alpha^k S^k$.

Les sommes partielles de la série sont donc des polynômes en S : ils commutent donc avec n'importe quel polynôme en S, et la limite également. En particulier, comme $I-S$ est un polynôme en S, il commute avec $(I-\alpha S)^{-1}$.

$$\text{Donc } \frac{d\pi(\alpha)}{d\alpha} = -{}^t v \cdot (I-S) \cdot (I-\alpha S)^{-2}.$$

On a alors : $\lim_{\alpha\to 0} \frac{d\pi(\alpha)}{d\alpha} = \lim_{\alpha\to 0} -{}^t v \cdot (I-S) \cdot (I-\alpha S)^{-2}$

$= -{}^t v \cdot (I-S) \cdot \lim_{\alpha\to 0}(I-\alpha S)^{-2} = {}^t v \cdot (I-S)$.

$$\lim_{\alpha\to 0} \frac{d\pi(\alpha)}{d\alpha} = {}^t v \cdot (I-S).$$

Etudions à présent la limite en 1. Pour cela, montrons d'abord que $(I-S)$ appartient à un groupe multiplicatif de matrices, en montrant que $Im(I-S) \oplus Ker(I-S) = \mathbb{R}^N$.
On a $||{}^t S||_\infty = \max_{1\le j\le N} \left(\sum_{k=1}^{n} |s'_{ij}|\right) = 1$, si ${}^t S = (s'_{ij})$,
car ${}^t S$ est stochastique. $(I - {}^t S)$ appartient donc, d'après le théorème 2, puis le théorème 1 de l'appendice, à un groupe multiplicatif.
Donc ${}^t(I - {}^t S) = I - S$ également.

Introduisons alors les matrices $Y = (I-S) \cdot (I-\alpha S)^{-2}$ et
$Z = (I-S)^\sharp \cdot (I-\alpha S)^2$. On a
$YZY = (I-S) \cdot (I-\alpha S)^{-2} \cdot (I-S)^\sharp \cdot (I-\alpha S)^2 \cdot (I-S) \cdot (I-\alpha S)^{-2}$
$= Y$ car $(I-S)$ et $(I-\alpha S)^{-2}$ commutent.
De même, $ZYZ = Z$, et $YZ = ZY$.
Donc $Y$ et $Z$ sont inverses pour le groupe de matrices considéré.

Donc pour $\alpha < 1$ :
$Z^\sharp = [(I-\alpha S)^2]^\sharp \cdot (I-S)^{\sharp\sharp} = (I-\alpha S)^{-2} \cdot (I-S)$, le $\sharp$ étant l'inverse si la matrice est inversible, on obtient donc finalement :

$$Z^\sharp = Y$$

Et si $\alpha = 1$, $Z = (I - S)^\sharp \cdot (I - S)^2 = E \cdot (I - S)$,
où E est l'élément neutre de groupe de matrices.

$$\text{Donc } Z^\sharp = (I - S)^\sharp E^\sharp = (I - S)^\sharp.$$

Puis $\lim_{\alpha \to 1}(I - S) \cdot (I - \alpha S)^{-2} = \lim_{\alpha \to 1} Y(\alpha) = \lim_{\alpha \to 1}(Z^\sharp(\alpha))$
$= \lim_{\alpha \to 1}(Z(\alpha))^\sharp = ((I - S)^\sharp \cdot (I - S)^2)^\sharp = (I - S)^\sharp$.

$$\text{Finalement } \lim_{\alpha \to 1} \frac{d\pi(\alpha)}{d\alpha} = {}^t v \cdot (I - S)^\sharp.$$

Cette dernière limite complète l'étude des variations du PageRank par rapport à $\alpha$. Nous avons vu que quand $\alpha$ tend vers 0, le Pagerank est très peu sensible ; c'est le contraire quand $\alpha$ tend vers 1. En effet, on sait que

$$(I - \alpha S)^{-1} = \sum_{k=0}^{N} \alpha^k S^k,$$

et quand $\alpha$ tend vers 1, cette série converge de moins en moins, et tend même à devenir divergente. Or $(I - \alpha S)^{-1} \underset{\alpha \to 1}{\longrightarrow} (I - S)^\sharp$, ce qui montre que quand $\alpha$ tend vers 1, la dérivée $|\frac{d\pi(\alpha)}{d\alpha}|$ est très forte.

La variation du PageRank semble donc beaucoup dépendre de l'ordre de grandeur de $\alpha$, en particulier de sa proximité aux bornes de son intervalle de variation ]0,1[. En effet :
- lorsque $\alpha$ est proche de 0, le PageRank est très insensible aux variations de $\alpha$.
- lorsque $\alpha$ se rapproche de 1, $|\frac{d\pi(\alpha)}{d\alpha}|$ devient de plus en plus important, ce qui signifie que le PageRank est très sensible aux variations de $\alpha$.

## 5.4. **Quelle valeur de $\alpha$ choisir ?**

$\alpha$ a donc un rôle très important pour les nombreux problèmes intervenant avec le PageRank, qu'il s'agisse de la vitesse de convergence ou des variations du vecteur. La valeur de $\alpha$ ne peut pas être prise au hasard, sans quoi Google n'aurait plus aucun sens.

Les faibles valeurs de $\alpha$ semblent, d'après les calculs précédents, être préférables aux fortes valeurs. En effet, lorsque $\alpha$ est faible, plusieurs

avantages apparaissent. Tout d'abord, la vitesse de convergence est élevée, ce qui garantit plus de facilité dans les calculs, et cela est primordial dans le traitement des milliards de pages du Web. Ensuite, plus $\alpha$ est proche de 0, moins le PageRank est sensible à ses variations. Cela est également très important, car le PageRank doit varier le moins possible si l'on veut que la méthode ait un sens.

Pourquoi alors Brin et Page ont-ils choisi $\alpha = 0.85$ ? Cela est dû à la signification première de ce paramètre. En effet, il pondère la part d'intervention des liens du Web par rapport au passage aléatoire d'une page à une autre. Plus $\alpha$ est proche de 1, plus la matrice hyperliens S est privilégiée, autrement dit, plus la structure du Web est présente dans la matrice Google. Par opposition, plus $\alpha$ est proche de 0, plus la matrice Google dépendra d'une distribution aléatoire de probabilités répartie sur l'ensemble des pages. Ainsi, comme le but de Google est de classer les pages par ordre d'importance, il semble logique de vouloir privilégier la matrice S représentant le graphe du Web. $\alpha$ doit donc être choisi très proche de 1 pour que le PageRank soit réaliste.

Ainsi, la valeur de $\alpha$ est très difficile à choisir. Google utilise actuellement $\alpha = 0.85$, ce qui semble être un bon arrangement entre la nécessité des valeurs fortes et celle des valeurs faibles. Cette valeur de $\alpha$ a effectivement été choisie par Brin et Page après de multiples essais, pour finalement se retrouver comme étant l'une des clés de la réussite planétaire du plus connu des moteurs de recherche.

## 6. Calcul du PageRank par la méthode de la puissance

Si le principe général de classement des pages a été expliqué dans les chapitres précédents, ainsi que les divers éléments entrant en jeu dans le choix des paramètres, notamment $\alpha$, ces problèmes dans la création de Google ne furent certainement pas les seuls qui se posèrent pour ses fondateurs. En effet, après que la théorie eût été mise en place, il restait pour eux l'essentiel du travail, c'est-à-dire la partie pratique, et l'application de la méthode. En particulier, l'un des principaux défis était celui de calculer ce fameux vecteur PageRank, point central du fonctionnement, pour ces dizaines de milliards de pages.

Aujourd'hui, les outils de calcul de ce vecteur sont inconnus du grand public. On imagine facilement des centaines d'ordinateurs surpuissants calculant vingt-quatre heures sur vingt-quatre, mais Google garde ces moyens secrets. Cependant, la méthode générale et l'algorithme utilisé dans les calculs sont publics. Il s'agit de la méthode dite de la puissance, bien connue dans le domaine des chaînes de Markov. Comment donc Google calcule-t-il, de manière simple, le vecteur invariant de classement des pages ?

### 6.1. **La méthode de la puissance : description et avantages.**

La méthode de la puissance est une méthode de calcul du vecteur invariant intervenant dans une chaîne de Markov. Il s'agit d'une méthode itérative, qui consiste à approcher à chaque étape le vecteur invariant. Dans le cadre de Google, chaque étape consistera en la multiplication matricielle de la matrice $G$ par le vecteur de classement des pages. Une telle méthode semble convenir : en effet, les coefficients de la colonne $j$ de la matrice $G$ correspondent à l'importance des liens pointant vers la page $j$. Plus cette colonne sera fournie, plus la multiplication à gauche par un vecteur-ligne modifiera ce dernier de manière à rendre son $j$-ième coefficient plus élevé. Si l'on itère ce calcul, le vecteur PageRank sera modifié de sorte que ses coefficients seront en phase avec l'importance de la page correspondante. C'est pourquoi si ces itérations convergent vers l'unique vecteur invariant (dont l'existence a été prouvée au chapitre 4), il s'agira bien du vecteur PageRank recherché. Nous verrons dans le paragraphe suivant ce qu'il en est de la convergence de la méthode de la puissance, dont nous allons rappeler le principe général :

On part pour cela d'un vecteur $q^0$, non nul, et à chaque étape, on réalise l'itération suivante :

$$q^k = r^{k-1} \cdot G$$

où $r^{k-1}$ est le vecteur $q^{k-1}$, à ceci près qu'il est normé.

En fait, normer les vecteurs est inutile dans le cadre de cette méthode, du moins à partir de la troisième itération, car ils le seront automatiquement. En effet :

$||q^2||_1 = q^2 \cdot e = r^1 \cdot G \cdot e = r^1 \cdot e$ car G est stochastique
d'où $||q^2||_1 = 1$ car le premier vecteur a été normé.

Finalement, les vecteurs sont automatiquement normés à partir de la troisième itération, ce qui rend superflu une telle rectification.

Pour résumer, la formule fondamentale du calcul du PageRank est $\pi = \pi \cdot (\alpha S + (1-\alpha)E)$, et $\pi \cdot {}^t(1 \;\; 1 \;... \; 1 \;\; 1) = 1$ car le PageRank doit être un vecteur de probabilités.

Cette méthode semble donc très simple, car il s'agit d'un simple procédé itératif, classique, et apparemment pratique à utiliser. Cependant, il existe de nombreuses autres méthodes de calcul du vecteur invariant d'une chaîne de Markov. Par exemple, nous avons vu que les puissances successives de la matrice $G$ tendaient vers la matrice dont les vecteurs-lignes sont égaux au vecteur invariant. De plus, la méthode de la puissance est connue pour la lenteur de son exécution, et il peut paraître étrange que ce soit celle à laquelle aient pensé Brin et Page dans leurs plans de calcul du PageRank, d'autant que la rapidité du calcul est un critère essentiel dans le fonctionnement de Google. Pourquoi donc ce choix ?

La méthode de la puissance possède en fait beaucoup d'avantages, en particulier dans le cadre de la matrice Google qui intervient. Tout d'abord, nous l'avons dit, c'est une méthode qui ne requiert que peu de calcul, et pour laquelle chaque itération est simple à réaliser. Il ne faut pas négliger ce genre d'avantages, car une trop grande complexité serait trop difficile à gérer, notamment à cause de la taille du vecteur concerné, et du nombre de pages entrant en jeu. Mais les calculs sont également facilités par la forme de la matrice de Google. En effet, développons l'expression traduisant l'invariance de $\pi$ par $G$ :

$$\pi^{(k+1)} = \pi^{(k)} \cdot G = \alpha\pi^{(k)} \cdot S + \frac{1-\alpha}{N}\pi^{(k)} \cdot e \cdot {}^t e$$
$$= \alpha\pi^{(k)} \cdot H + (\alpha\pi^{(k)} \cdot a + 1 - \alpha) \cdot {}^t e \frac{1}{N}.$$

Nous nous apercevons ainsi que l'itération porte essentiellement sur la matrice $H$ (introduite au chapitre 2). C'est l'une des raisons principales pour lesquelles la méthode de la puissance a été privilégiée. En effet, la matrice H est une matrice extrêmement creuse, car il est finalement assez rare qu'une page pointe vers une autre. Ainsi, la forte présence de zéros facilitera l'itération. De plus, ces itérations ne font pas intervenir la matrice $G$, matrice qui est, rappelons-le, totalement pleine. D'autres méthodes de calcul ne pourraient s'effectuer uniquement que sur la matrice $G$, ce qui rendrait les calculs encore plus lourds et difficiles à réaliser. En particulier, la méthode qui consisterait à calculer la limite de $G^n$ serait beaucoup trop lourde à gérer, car en fait d'un vecteur et d'une matrice creuse, il s'agit d'effectuer des opérations sur une matrice de $10^{20}$ coefficients, totalement pleine !

Un autre avantage de cette méthode de la puissance concerne la mémoire informatique nécessaire au calcul du PageRank. En effet, il est préférable, informatiquement parlant, de traiter le moins de données possibles, pour fluidifier les opérations. La méthode de la puissance nécessite uniquement de connaître la matrice $H$ (ce qui n'est pas très contraignant, car elle est très creuse), et le vecteur PageRank à chaque itération. Cela permet donc de faciliter les calculs et d'arriver en une complexité moindre au résultat attendu. Ceci n'est d'ailleurs pas une économie négligeable, car rappelons que les tailles des matrices et des vecteurs sont de l'ordre de $10^{10}$ !

La méthode de la puissance possède également un autre avantage par rapport aux autres méthodes. En effet, chaque itération est très simple, et nécessite uniquement un calcul matriciel élémentaire. D'autres méthodes, certes plus directes, demanderaient de modifier à chaque étape les coefficients de la matrice $G$ pour déterminer le vecteur invariant. Voilà pourquoi les méthodes itératives, dans le cas de gigantesques matrices, sont privilégiées.

La méthode de la puissance a donc été sélectionnée par Brin et Page parmi beaucoup d'autres méthodes pour les nombreuses qualités qu'elle possède. Elle peut certes paraître à première vue assez rudimentaire,

par sa simplicité, mais est finalement celle qui permet d'effectuer des calculs le plus facilement possible, car elle évite de nombreux écueils dus à la complexité, notamment informatique, qu'engendre le traitement d'autant de données, de matrices, et de vecteurs.

### 6.2. **Vitesse de convergence de la méthode de la puissance.**

Si la méthode de la puissance possède autant de qualités, il n'en reste pas moins qu'elle se doit d'être efficace au niveau du calcul. En effet, il faut absolument que l'algorithme de calcul converge rapidement vers le vecteur invariant cherché, sans quoi il restera fatalement vain. Nous avons vu au chapitre 5 que c'est $\alpha$ qui régissait cette vitesse de convergence. Néanmoins, nous avions besoin pour montrer cela de montrer que la vitesse de convergence dépend uniquement de la vitesse à laquelle $|\lambda_2|^k \to 0$, où $\lambda_2$ est la deuxième valeur propre de $G$ (par ordre décroissant des modules). Ce paragraphe s'applique à démontrer ce résultat, avec, au passage, la démonstration la plus importante concernant le calcul du PageRank, celle sans laquelle la recette serait plus que caduque car totalement inefficace : il est temps de montrer que la méthode de la puissance converge effectivement vers le vecteur PageRank !

Nous ne traiterons ici que le cas où $G$ est diagonalisable.
Dans ce cas, on a $Sp(G) = \{1, \lambda_2, ..., \lambda_p\}$, et on sait que
$|\lambda_1| = 1 > |\lambda_2| > ... > |\lambda_p|$.
Soit $P_k$ le projecteur spectral sur $E_{\lambda_k}(G)$ parallèlement à $\underset{j\neq k}{\oplus} E_{\lambda_j}(G)$.

On a alors $G = \lambda_1 P_1 + ...\lambda_p P_p$. En effet, soit $X \in \mathbb{R}^{\mathbb{N}}$.
On a $X = a_1 e_1 + ... + a_p e_p$ où $e_i \in E_{\lambda_i}(G)$ car les $E_{\lambda_i}(G)$ sont supplémentaires.

Donc $G \cdot X = a_1 G \cdot e_1 + ... + a_p G \cdot e_p$
$= a_1\lambda_1 e_1 + ... + a_p\lambda_p e_p$
$= \lambda_1 P_1(X) + ... + \lambda_p P_p(X)$,

$$\text{D'où } G = \lambda_1 P_1 + ...\lambda_p P_p.$$

Notons alors $q^k$ le vecteur PageRank à la k-ième itération, en considérant tous les vecteurs $q^k$ normés.

On a $q^k = q^0 G^k = q^0 P_1 + \lambda_2^k q^0 P_2 + ... + \lambda_p^k q^0 P_p$.
D'où $q^k - P_1 q^0 = \lambda_2^k q^0 P_2 + ... + \lambda_p^k q^0 P_p$ ,

c'est à dire : $||q^k - q^0 P_1||_1 \leq |\lambda_2^k| \cdot ||q^0 P_2||_1 + ... + |\lambda_p^k| \cdot ||q^0 P - p||_1$
$\leq |\lambda_2|^k (|q^0 P_2|||_1 + ... + ||q^0 P_p||_1)$ car $\lambda_2 \geq \lambda_i \ \forall i \geq 2$.
Or $|\lambda_2| < 1$, donc $||q^k - q^0 P_1||_1 \underset{k \to +\infty}{\longrightarrow} 0$, soit $q^k \underset{k \to +\infty}{\longrightarrow} q^0 P_1$,

et ceci à la même vitesse de convergence que $|\lambda_2|^k$.

Il faut alors montrer que $q^0 P_1 = \pi$.
On sait que $q^0 G \cdot P_1 = q^0 P_1 \cdot G$ car $G \cdot P_1 = P_1 \cdot G$.
D'où $q^0 G \cdot P_1 = q^0 P_1 \cdot ((P_1 + \lambda_2 P_2 + ... + \lambda_p P_p)) = q^0 P_1^2$
car $P_1 \cdot P_i = 0 \ \forall \ i \neq 1$ et comme $P_1$ est un projecteur, on en déduit que $q^0 G \cdot P_1 = q^0 P_1$, i.e. $q^0 P_1 \cdot G = q^0 P_1$.

Donc $q^0 \cdot P_1$ est un invariant pour la multiplication à droite par $G$, c'est à dire que $\exists \ \lambda \in \mathbb{R}$, $q^0 \cdot P_1 = \lambda \pi$, d'où $q^k \underset{k \to +\infty}{\longrightarrow} \lambda \pi$. Or $||q^k||_1 = 1$, donc $||\lambda \pi||_1 = 1$,
d'où $|\lambda| \times ||\pi||_1 = 1$, soit $|\lambda| = 1$. Finalement $q^0 \cdot P_1 = \pi$ (car la forme de $P_1$ contraint $\lambda$ à être réel et strictement positif). Donc

$$q^k \underset{k \to +\infty}{\longrightarrow} \pi.$$

et ce avec la vitesse de convergence de $|\lambda_2|^k$.

La décomposition en sous-espaces propres n'est toutefois plus valable pour une matrice non diagonalisable, on s'appuie donc sur le théorème de décomposition des noyaux pour obtenir une majoration similaire.

C'est donc $\lambda_2$ qui donne la vitesse de convergence de la méthode de la puissance, et comme nous l'avons vu en partie 5, cela ramène à montrer que c'est $\alpha$ qui gouverne la convergence.

Ainsi la méthode de la puissance permet-elle d'effectuer les calculs avec la complexité informatique la plus faible, mais également d'être maître de la vitesse de convergence. Avec $\alpha = 0.85$, comme il est d'usage de le choisir ainsi, on approche au bout d'une cinquantaine d'itérations d'un vecteur certes pas tout à fait exact, mais tout à fait satisfaisant compte tenu du but recherché et des paramètres qui entrent en jeu.

La méthode de la puissance permet donc une vitesse de convergence assez élevée, ce qui est l'un des points clés de la détermination du vecteur PageRank. Elle semble donc tout particulièrement adaptée au calcul de ce vecteur, et c'est la raison pour laquelle elle a surpassé tout

autre type de méthode, plus récentes, plus modernes, de détermination du vecteur invariant d'une chaîne de Markov.

### 6.3. **Algorithmique et programmation de la méthode de la puissance.**

Nous l'avons vu, la méthode de la puissance semble être d'une simplicité surprenante. Il s'agit d'un court calcul itératif, à la portée du premier informaticien en herbe venu. Cette partie s'appliquera à déterminer un programme permettant de calculer le PageRank d'un graphe entre pages. Nous effectuerons la programmation dans le langage Maple, qui certes est souvent privilégié pour son calcul formel, mais permet également de confectionner des programmes légers, rapides et efficaces (on utilisera la commande **with (LinearAlgebra)**).

Le but des programmes suivants est de donner les résultats cherchés à partir de la matrice la plus simple entrant dans la construction de la matrice $G$, celle contenant un 1 lorsqu'une page pointe vers une autre, et un 0 sinon. De cette matrice, nous déterminerons les matrices $H$, $S$, et $G$, puis enfin pourrons calculer le fameux PageRank correspondant.

Commençons par écrire un programme donnant la matrice $S$ à partir de la matrice des liens. Il s'agit simplement de diviser une ligne par la somme des coefficients lorsque celle-ci est non nulle (car $S$ est stochastique en ligne), et lorsque la ligne ne contient que des 0, de la remplir de coefficients égaux à $\frac{1}{n}$. On y arrive avec une simple boucle 'for' :

```
somme :=proc(M,i) local j,z ; z :=0 ;
for j from 1 to ColumnDimension(M) do
z :=(M[i,j]+z) ; od ; z ; end ;
```

(renvoie la somme des coefficients d'une ligne d'une matrice)
Puis

```
mat2 :=proc(H) local i,j,m,n,z ;
m :=RowDimension(H) ; n :=ColumnDimension(H) ;
for i from 1 to m do z :=(somme(H,i)) ;
if z≠0 then
for j from 1 to n do H[i,j] :=(H[i,j]/z) ; od ;
else for j from 1 to n do H[i,j] :=(1/n) ;
od ; fi ; od ; H ; end ;
```

(Attention cependant : la matrice $H$ que l'on entre n'est pas exactement la même que celle définie au chapitre 3 : dans le cas présent, les lignes non nulles n'ont pas encore été divisées par le nombre de liens vers laquelle pointe la page correspondante)

Cette procédure permet de construire $S$ à partir de la matrice du Web. Construisons à présent $G$, à partir de $S$. Pour cela, il suffit de pondérer $S$ par le coefficient $\alpha$ ('a' dans le programme suivant), et d'ajouter la matrice $E$ pondérée par le coefficient $1-\alpha$. Dans le programme suivant, l'utilisateur pourra également choisir la valeur de $\alpha$ souhaitée, ce qui permettra d'étudier l'influence de ce paramètre.

```
matG :=proc(S,a) local G,E;
E :=Matrix(RowDimension(S),ColumnDimension(S),
1/RowDimension(S));
G :=a*S + (1-a)*E; G; end;
```

Cette procédure construit ainsi $G$ à partir de $S$ et du coefficient $\alpha$ choisi. Il ne reste plus qu'à combiner les deux procédures pour que le programme renvoie la matrice $G$, à partir de la matrice de liens, et du paramètre $\alpha$ donné par l'utilisateur. Le programme final est :

```
constructor := proc (H,a) local G;
G :=matG(mat2(H),a); G; end;
```

La matrice $G$ est donc plutôt aisée à construire à partir d'un graphe Web donné. Il suffit pour cela de construire une matrice carrée dont la taille est le nombre de pages ; de mettre un 1 en $(i, j)$ lorsque la page $i$ pointe vers la page $j$ ; de la compléter avec des zéros ; et d'appliquer le programme précédent. On obtient alors la matrice $G$, telle qu'elle fut élaborée par Brin et Page, et qui permet, par la méthode de la puissance, de déterminer le PageRank qui ordonne les pages par ordre d'importance.

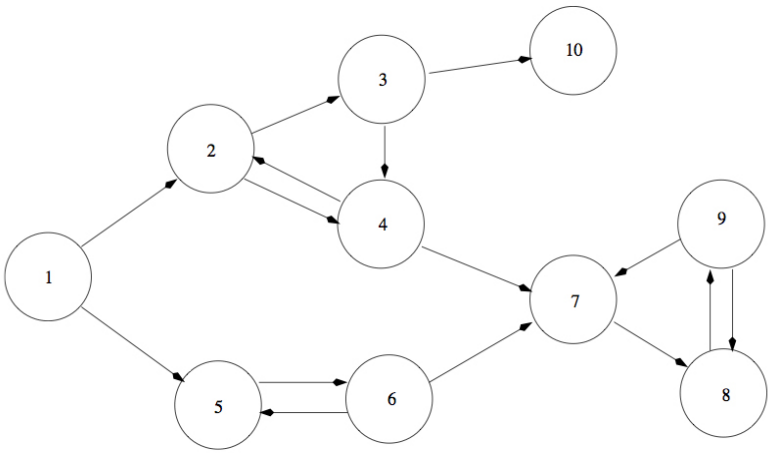
Nous pouvons maintenant mettre en place le programme de méthode de la puissance en lui-même, qui, à partir de la matrice $G$, donnera les coefficients d'importance de chaque page. L'utilisateur du programme entrera comme paramètre la matrice G, un vecteur de départ $V$ de la méthode de la puissance, ainsi que le nombre $n$ d'itérations qu'il souhaite effectuer. La méthode ayant été expliquée au début de ce chapitre, nous pouvons donner le programme correspondant :

```
puiss :=proc(V,G,n) local i,Z ; Z :=V ;
for i from 1 to n do
Z :=VectorMatrixMultiply(Z,G) ; od ; Z ; end ;
```

Cette procédure effectuera ainsi les $n$ itérations nécessaires, pour donner le vecteur PageRank invariant par G. Cependant, le vecteur solution ne sera pas forcément normé, car rien ne l'indique dans le programme, mais il ne s'agit que d'un problème mineur car la dimension du sous-espace vectoriel propre associé à 1 est précisément 1. C'est-à-dire que deux vecteurs invariants par $G$ seront colinéaires. Il suffira donc, pour obtenir le véritable vecteur PageRank comme distribution de probabilités, de normer le vecteur obtenu si ce n'est pas déjà fait, en divisant chacune de ses composantes par la racine de la somme des composantes. On voit alors arriver l'unique, et le tant attendu vecteur PageRank, pilier de l'originalité de Google. Ce programme permet de se rendre compte de la simplicité de la méthode de la puissance, par sa longueur très réduite, et la commodité de l'algorithmique mise en jeu. Il est cependant évident que les calculateurs de Google n'utilisent pas un programme aussi simple, car il serait impossible de gérer des milliards de pages, mais l'idée est là. Et si Google tient à garder la recette du calcul secrète, cette simplification peut néanmoins être utilisée dans le traitement de mini-webs de quelques pages tout au plus, qui, sans représenter une parfaite réalité, permettent de tester les grandes lignes du principe général de gestion d'un graphe comme Internet.

## 6.4. **Calcul du PageRank sur un graphe de 10 pages.**

Pour clore ce chapitre, nous proposons de tester les procédures précédentes sur un exemple. Nous travaillerons sur le mini-web de 10 pages ci-dessous.




Rappelons encore une fois que ce genre d'exemple n'a rien à voir avec la réalité des $10^{10}$ pages du Web, mais il est intéressant de voir comment peuvent s'appliquer les programmes, même si cela reste très schématisé.

Commençons, à partir de ce graphe, par construire la matrice la plus simple le représentant. Il s'agit de la matrice indiquant simplement les liens, c'est-à-dire qui contient un 1 en position $(i, j)$ si la page $i$ pointe vers la page $j$. Cette matrice est donnée par

$$\mathrm{A} := \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A partir de cette matrice, le graphe du mini-web est donc parfaitement déterminé, et tous les calculs peuvent être effectués. Commençons par donner la matrice $S$ correspondante :

**S :=mat2(A) ;**

$$\mathrm{S} := \begin{pmatrix} 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10 \end{pmatrix}$$

Il faut à présent, pour que la matrice soit complète, rajouter la correction en $(1-\alpha)E$ relative à l'éventuelle téléportation d'un surfer aléatoire On choisira ici $\alpha = 0.85$. On obtient alors la matrice $G$ :

**G :=constructor(A,0.85) ;**

$$\mathbf{G} := \begin{pmatrix} 0.015 & 0.44 & 0.015 & 0.015 & 0.44 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\ 0.015 & 0.015 & 0.44 & 0.44 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 \\ 0.015 & 0.015 & 0.015 & 0.44 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.44 \\ 0.015 & 0.44 & 0.015 & 0.015 & 0.015 & 0.015 & 0.44 & 0.015 & 0.015 & 0.015 \\ 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.865 & 0.015 & 0.015 & 0.015 & 0.015 \\ 0.015 & 0.015 & 0.015 & 0.015 & 0.44 & 0.015 & 0.44 & 0.015 & 0.015 & 0.015 \\ 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.865 & 0.015 & 0.015 \\ 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.865 & 0.015 \\ 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.015 & 0.44 & 0.44 & 0.015 & 0.015 \\ 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 & 0.10 \end{pmatrix}$$

La matrice $G$ de ce mini-web est donc maintenant construite. On remarque que conformément aux étapes de construction, cette matrice est stochastique et à coefficients strictement positifs, ce qui, d'après le chapitre 4, assure l'existence et l'unicité du vecteur invariant. Nous avons de plus vu précédemment que, de ce fait, la méthode de la puissance était certaine de converger vers ce point fixe. Appliquons alors le programme de la méthode de la puissance pour obtenir le PageRank

relatif à chacune des dix pages :

**V :=**
**Vector[row]([1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10,1/10]) :**
**puiss(V,G,50) ;**

$$\mathbf{V :=}$$
$$\begin{pmatrix} 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \end{pmatrix}$$

Ainsi, il est possible de déterminer le point fixe de la chaîne de Markov relative à ce graphe par le simple procédé itératif de la méthode de la puissance. Cependant, nous avons vu au chapitre 4 que la matrice $G$ avait la propriété suivante : $G^k \underset{k\to+\infty}{\longrightarrow} G_\infty$, où $G_\infty$ est la matrice dont tous les vecteurs-lignes sont égaux au vecteur invariant. Nous avons remarqué précédemment que ce genre de calcul était plus lourd et moins intéressant que la méthode de la puissance pour déterminer le vecteur invariant ; néanmoins, il serait intéressant de programmer ce résultat, et de tester, avec le graphe précédent, cette propriété. La procédure suivante donne, en prenant comme paramètre la matrice $G$, et un entier $n$, la matrice $G^n$. En prenant un $n$ assez grand, on devra arriver à une matrice dont les vecteurs-lignes sont sensiblement égaux au vecteur invariant.

**puissanceG := proc(G,n) local Z,i ;**
**Z :=G ; for i from 1 to n do**
**Z :=MatrixMatrixMultiply(G,Z) ; od ; Z ; end ;**

Appliquons maintenant ce programme à la matrice G relative au mini-web précédent :

**puissanceG (G,50) ;**

$$\begin{pmatrix} 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \\ 0,0179 & 0,049 & 0,039 & 0,055 & 0,052 & 0,062 & 0,173 & 0,270 & 0,248 & 0,034 \end{pmatrix}$$

La matrice obtenue a pour vecteur-lignes le vecteur invariant trouvé précédemment. Les deux méthodes semblent donc assez efficaces pour déterminer le vecteur PageRank, même s'il est évident que la seconde est bien trop délicate à mettre en œuvre pour un réseau de $10^{10}$ pages !

L'exemple du mini-web précédent paraît finalement être concluant dans la recherche du vecteur invariant, par le fait que les deux méthodes mises en œuvre donnent un résultat identique, cohérent à ce que l'on pouvait attendre au vu de la structure générale du graphe. Il est vrai que cela peut paraître dérisoire de traiter des graphes si simplifiés, car Internet est un milliard de fois plus complexe aujourd'hui ! Mais pourtant, même les géniaux Sergey Brin et Larry Page ont dû passer par là pour mettre en place l'outil de recherche le plus utilisé de nos jours ...

## 7. Appendice sur les groupes multiplicatifs de matrices

***Définition :***

On appelle groupe multiplicatif tout ensemble $\mathcal{G}$ de matrices carrées tel que $(\mathcal{G}, \times)$ soit un groupe. En particulier, on notera que :
• $\exists\ E \in \mathcal{G},\ \forall\ Y \in \mathcal{G},\ E \cdot Y = Y \cdot E = Y$
• $\forall\ Y \in \mathcal{G},\ \exists\ Z \in G,\ Y \cdot Z = Z \cdot Y = E$

***Proposition 1 :***

E est un projecteur de $\mathcal{M}_n(\mathbb{K})$. En effet, $E^2 = E$ car $E$ est élément neutre. En particulier, on a $Ker(E) \oplus Im(E) = \mathbb{K}^n$.

***Proposition 2 :***

$\forall\ Y \in \mathcal{G}$, rg$(Y)$ =rg$(E)$.

**Preuve** :

$Y = E \cdot Y$ donc $Im(Y) \subset Im(E)$ et de plus,
$\exists\ Z \in \mathcal{G}, Y \cdot Z = E$ donc $Im(E) \subset Im(Y)$, c'est à dire rg$(E)$ ≤rg$(Y)$ d'où rg$(Y)$ =rg$(E)$, $\forall\ Y \in \mathcal{G}$.

***Proposition 3 :***

Soit $Y \in \mathcal{G}$. On a $Y = P \cdot \begin{pmatrix} Y' & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$ avec $P \in GL_n(\mathbb{K})$ et $Y' \in GL_r(\mathbb{K})$, $r$ =rg$(E)$ =rg$(Y)$.

**Preuve** :

$E$ est un projecteur. $\exists\ P \in GL_n(\mathbb{K}),\ E = P \cdot \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$. Or $Y \cdot E = E \cdot Y = Y$, donc $Ker(E) = Ker(Y)$ et $Im(E) = Im(Y)$ donc dans la base de $Ker(E) \oplus Im(E)$, on a $Y = \begin{pmatrix} Y' & 0 \\ 0 & 0 \end{pmatrix}$.

Finalement, P étant la matrice de passage de la base de $Ker(E) \oplus Im(E)$ à celle considérée, on a $Y = P \cdot \begin{pmatrix} Y' & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$.

**Remarque** : On notera que l'on a alors $Z = P \cdot \begin{pmatrix} Y'^{-1} & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$.

***Théorème 1 :***

Il est équivalent de dire :
(i) $\exists\ \mathcal{G}$, groupe multiplicatif tel que, $Y \in \mathcal{G}$.
(ii) $Ker(Y) \oplus Im(I) = \mathbb{K}^n$.

**Démonstration :**

(i) $\Rightarrow$ (ii) : $Ker(Y) = Ker(E)$ et $Im(Y) = Im(E)$
donc $Ker(Y) \oplus Im(Y) = \mathbb{K}^n$
(ii) $\Rightarrow$ (i) : Soit $P$ le projecteur sur $Im(Y)$ parallèlement à $Ker(Y)$. Soit $P$, matrice de passage de la base canonique de $\mathbb{K}^n$ à une base de $Ker(Y) \oplus Im(Y)$.
$E = P \cdot \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$ comme projecteur, et $Y = P \cdot \begin{pmatrix} Y' & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$, $Y \in GL_n(\mathbb{K})$.
Soit $\mathcal{G} = \{P \cdot \begin{pmatrix} U & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}, U \in GL_n(\mathbb{K})\}$.

On a :
• $\forall\ Y \in\ \mathcal{G},\ E \cdot Y = Y \cdot\ E = Y$.
• $\forall\ Y \in\ \mathcal{G},\ \exists\ Z\ \in \mathcal{G},\ Y \cdot Z = Z \cdot\ Y = E$.
(On pose $Z = P \cdot \begin{pmatrix} Y'^{-1} & 0 \\ 0 & 0 \end{pmatrix} \cdot P^{-1}$)
• L'associativité est héritée de celle sur $\mathcal{M}_n(\mathbb{K})$.
• $\times$ est une loi interne sur $\mathcal{G}$.
$\mathcal{G}$ est bien un groupe multiplicatif de matrices.

***Théorème 2 :***

Soit $u \in \mathcal{L}(E)$, tel que $||u|| \leq 1$, $||\ ||$ étant une norme multiplicative. Alors $Ker(Id - u) \oplus Im(Id - u) = E$.

**Démonstration :**

Soit $v_k \in \mathcal{L}(E)$, $v_k = \dfrac{Id + u + ... + u^{k-1}}{k}$.
Montrons que $Ker(Id - u) \cap Im(Id - u) = \{O_E\}$.

• Soit $x \in Ker(Id - u) \cap Im(Id - u)$. On a $x = u(x)$,
alors pour $k \to +\infty$, on a $v_k(x) = x \to x$.

• Soit $x \in Im(Id - u)$. $\exists\ t \in E,\ x = t - u(t)$.
Alors $v_k(x) = \dfrac{t - u(t) + u(t) - u^2(t) + ... + u^{k-1}(t) - u^k(t)}{k} = \dfrac{t - u^k(t)}{k}$.
D'où $||v_k(x)|| \leq \dfrac{||t|| + ||u^k(t)||}{k} \leq \dfrac{||t|| + |||u^k||| \times ||t||}{k} \leq \dfrac{||t|| + |||u|||^k||t||}{k}$

et comme $||u|| \leq 1$, alors $|||u||| \leq 1$ (avec $|||u||| = \sup\limits_{||t||=1} ||u(t)||$).
Donc $||v_k(t)|| \leq \frac{2||t||)}{k} \to 0$ quand $k \to +\infty$.

• Ainsi, pour $k \to +\infty$, on a à la fois $v_k(x) \to x$ et $v_k(x) \to 0$. Donc $x = 0_E$. La somme $Ker(Id - u) \oplus Im(Id - u)$ est directe.

• On a de plus $\text{rg}(Id - u) + \dim(Ker(Id - u)) = \dim(E)$,
d'où $Ker(Id - u) \oplus Im(Id - u) = E$.

***Proposition 4 :***

Notons $A^\sharp$ l'inverse de $A$ pour le groupe multiplicatif considéré.
Il est équivalent de dire :
(1) $Y = Z^\sharp$
(2) $Y \cdot Z \cdot Y = Y, \quad Z \cdot Y \cdot Z = Z, \quad Y \cdot Z = Z \cdot Y$

***Proposition 5 :***

$$\lim_{\alpha\to 1^-} Z^\sharp(\alpha) = [\lim_{\alpha\to 1^-} Z(\alpha)]^\sharp.$$

**Preuve** :

On a $\forall a \in ]0,1[$ :

$Z^\sharp(\alpha)\cdot Z(\alpha)\cdot Z^\sharp(\alpha) = Z^\sharp(\alpha)$
$Z(\alpha)\cdot Z^\sharp(\alpha)\cdot Z(\alpha) = z(\alpha)$
$Z(\alpha)\cdot Z^\sharp(\alpha) = Z^\sharp(\alpha)\cdot Z(\alpha)$.

Faisons tendre $\alpha$ vers 1, posons $R = \lim_{\alpha\to 1^-} Z(\alpha)$, et $T = \lim_{\alpha\to 1^-} Z^\sharp(\alpha)$.

On a $T\cdot R\cdot T = T$
$R\cdot T\cdot R = R$
$T\cdot R = R\cdot T$.
Donc $T = R^\sharp$.

## 8. Conclusion

Altavista, Yahoo et les autres ne peuvent que s'incliner. Google est de nos jours le maître incontesté des moteurs de recherche. Il aura fallu moins de cinq ans pour que Google acquière le monopole des outils de recherche sur la Toile.

Derrière cette réussite se cache un nom moins connu, celui de PageRank, qui explique ce succès invraisemblable. Le PageRank, comme nous l'avons vu à travers cet ouvrage, est à la base de l'originalité du fonctionnement de Google. Personne auparavant n'avait eu cette idée de classer les pages en rapport avec leur importance dans le graphe du Web. Toujours est-il que Larry Page et Sergey Brin, par cette invention, ont révolutionné les moteurs de recherche et totalement éradiqué la concurrence.

Chacun de nous sait se servir de Google, mais qui en connaissait le principe ? Les mystères vous sont à présent connus (si du moins vous n'avez pas commencé par cette page), et vous pourrez, au prochain clic sur l'onglet Recherche Google que tout le monde connaît, vous dire qu'en un quart de seconde, les pages correspondant au sujet de votre recherche sont classées par ordre de PageRank décroissant, PageRank qui au préalable a été calculé par la méthode de la puissance. Lorsque vous créerez une nouvelle page Web, vous pourrez vous dire que vous rajoutez une ligne et une colonne à l'immense matrice Google, que vous modifiez le PageRank de toutes les pages de la Toile (de manière, rassurez-vous, totalement insignifiante) ...

Le PageRank est l'ingrédient secret qui rend la recette tellement surprenante. Cet ingrédient est toujours d'une simplicité inouïe, mais il faut pourtant une grande expérience pour le confectionner, pour prévoir les conséquences qu'il apportera dans la formule finale, et pour savoir comment ses qualités se marieront-elles avec celles des autres éléments entrant en jeu dans la composition ultime du mélange. Il faut enfin que cet ingrédient ne soit pas altéré par le contact avec les autres éléments et garde toutes ses propriétés, c'est-à-dire qu'il reste invariant lors du produit final issu de lui-même et du reste des composants de la recette. Et cette spécialité de la maison porte toujours le nom du chef qui l'a découverte : le PageRank, par Larry Page...

Un jour peut-être deux autres étudiants découvriront-ils une autre méthode, plus efficace, plus rapide, et plus simple pour classer les pages Web. Mais aujourd'hui, Google est intouchable. Et ses moyens de calcul restent eux aussi intouchables pour le grand public. Nous vous avons dévoilé une partie des mystères de Google, mais il reste une nappe de brouillard sur ce côté pratique du fonctionnement... L'énigme Google n'est pas entièrement résolue, et personne d'autre que ses fondateurs ne l'ont résolue. La recette du chef reste donc en partie secrète, et peut-être que personne ne saura jamais avec quels moyens elle aura pu surpasser, et ce de loin, n'importe quel autre moteur de recherche.

**Remerciements :**

Pour commencer, nous remercions Monsieur Combrouze pour l'aide à la compréhension des points délicats.

Mais nous pensons aussi à Monsieur Combrouze pour avoir pu se procurer les documents utiles à ce travail.

Enfin, pour son soutien et son humour, nous remercions également Monsieur Combrouze.

**Bibliographie :**

*Google's PageRank and beyond*, de A.N. Langville et C.D. Meyer, qui nous a fait travailler notre anglais.

Problème d'ESSEC 2008, épreuve de mathématiques n° 2, et Problème d'Agrégation externe de mathématiques 2008, qui nous ont fait réfléchir tout en en apprenant.

Et enfin www.Google.fr, sans quoi ce projet n'aurait assurément pas pu être mené.

*Les amis des mathématiques* vous recommandent également une nouveauté en exclusivité :
*L'algèbre du trafic routier*, par les trois brillants scientifiques G. Karam, P. Jeunesse et M. Viallon.