Une activité Google en collège : Tableur et probabilités

Françoise et Sébastien ESTRADE Collège Louise Michel Chagny et collège La Chataigneraie à Autun
francoise.estrade@ac-dijon.fr *sebastien.estrade@ac-dijon.fr*

Résumé : L'activité présentée ici (principalement en partie II), s'est déroulée dans une **classe de troisième** et a permis à des élèves de découvrir des **mathématiques cachées derrière le fonctionnement du moteur de recherche google**. Elle a également servi au professeur à faire manipuler le **tableur** à ses élèves et introduire une notion de **probabilité**.

Mots clés : Activité en collège ; tableur ; algorithme ; classement pages google ; algorithme PageRank ; probabilité

Ayant toujours cherché à montrer aux élèves l'utilisation des mathématiques dans les nouvelles technologies, le stage « Mathématiques et nouvelles technologies » animé par Catherine Labruère-Chazal et Denis Gardes nous a apporté des idées et des notions exploitables avec des élèves de collège.
Nous avons découvert à cette occasion l'algorithme PageRank du moteur de recherche Google qui est un algorithme permettant de classer les pages Web selon leur popularité. Il nous avait été présenté selon deux points de vue : les matrices ou le tableur. L'utilisation du tableur nous a semblé opportune pour introduire la notion de probabilité à des élèves de 3e.

1. Conditions de l'activité

- Journées à thèmes au Collège La Châtaigneraie à Autun pour des élèves de 3e.
- Durée totale : 3h en salle info avec 1 vidéoprojecteur pour le professeur et 1 ordinateur relié à internet par groupe de deux élèves.

2. Présentation de l'activité

- Recherche sur l'histoire de la société Google : les élèves ont un questionnaire (voir l'activité élèves, partie I) à remplir sur l'année de naissance, le nom des créateurs, les

activités de la société, ses revenus, etc. Ils ont comme outil de recherche internet et Wikipédia (principalement). Ce questionnaire, corrigé à l'aide d'un diaporama (résumé en fin d'activité élèves, partie I), permet d'introduire la multitude des activités et services proposés par la société Google.

• Activité sur l'algorithme PageRank : les élèves doivent remplir une fiche (voir activité élèves, partie II). Les calculs, simples au début, se répètent et deviennent plus fastidieux. Les élèves travaillent alors au tableur.

• On peut présenter d'autres exemples (de réseaux) avec leurs classements, voire même montrer la matrice servant à résoudre plus rapidement un exemple plus complexe. (voir à ce sujet l'annexe pour des explications plus détaillées destinées aux professeurs ou aux élèves de lycée un peu curieux).

• Le fait d'utiliser le tableur permet, de valider quelques items du B2I et aussi d'introduire les probabilités.

Nous n'avons parlé de probabilité qu'en fin d'activité. Par ailleurs, le fait d'avoir des fractions dans les calculs rajoutait une difficulté que nous voulions contourner. Nous avons donc choisi de présenter l'activité aux élèves avec des points de popularité (voir activité élève, partie II). Le tableur permet de changer aisément les conditions initiales et de constater que le résultat final de fréquentation des internautes sur une page ne dépend que des liens entre les pages. Les points de popularité permettent d'aboutir à une probabilité de présence d'un internaute sur une page Web à un instant donné (voir activité élèves, partie II).

3. Bilan et modifications à envisager

La séance s'est très bien déroulée avec une classe de troisième très intéressée. Cependant nous nous sommes aperçu de certains problèmes :

Nous avons toujours cherché à simplifier au maximum le travail sur PageRank étant persuadés que l'algorithme était complexe à saisir, que nous y arrivions parce que nous étions plongés dedans depuis quelques temps mais que pour de jeunes élèves qui devaient le comprendre en quelques minutes cela serait plus difficile. Nous avons donc passé beaucoup de temps à le leur détailler, et à le leur présenter à plusieurs reprises : un tableau, puis le tableur puis un diaporama (voir activité élèves, parties I et II).

Au final ils l'ont très bien compris et à l'avenir, nous supprimerons les phrases : « A reçoit la moitié des points de..... » pour leur laisser davantage d'autonomie dans leurs recherches. De plus nous avons regretté de ne pas avoir intégré sur leur feuille d'autres exemples à travailler au tableur (à cinq ou six pages Web). La modification sera faite pour les prochaines fois.

Nous avons même « joué » avec eux, sur un réseau à 8 pages à deviner quelle page obtiendrait la plus forte popularité connaissant le réseau. Généralement nous tombions d'accord et notre choix s'avérait juste une fois vérifié au tableur. (voir exemples en annexe)

Activité élèves : Partie I

Connaissez-vous Google ?

> *Vous connaissez tous le site www.google.fr pour effectuer des recherches sur internet mais connaissez-vous la société Google, ses créateurs, son origine et le fonctionnement du site Google.fr ?*

La société Google :

Lien :http://fr.wikipedia.org/wiki/Google

Qui sont les créateurs de la société Google ? En quelle année a-t-elle été créée ? Où ?
D'où vient le nom de cette société ? Que signifie-t-il ?
Quelle est l'origine des revenus de cette société ?

Les activités :

Lien : http://logiciels.zorgloob.com/liste.php

Parmi tous les sites (ou logiciels) créés (ou rachetés) par la société Google, citer les plus connus :

.....

.....

Lien : http://fr.wikipedia.org/wiki/Google#Critiques_et_controverses

> *Ce très grand nombre de sites, de services, et de logiciels disponibles jette également le doute sur les intentions de la société Google puisque qu'elle collecte des informations sur ses utilisateurs pour les stocker et les revendre.*

Le site Google.com (ou Google.fr pour la France)

Lien : http://fr.wikipedia.org/wiki/Google_(moteur_de_recherche)

Comment s'appelle le système de classement sur lequel est basé le site :

.....

Qu'indique ce système ?

.....

Ce système a-t-il été un succès ?

.....

Éléments de réponse :

Les créateurs de google sont Larry Page et Sergueï Brin. La société a été créée en 1998 en Californie, avec de petits moyens : ses débuts se sont faits dans un garage avec un matériel très réduit.
Le nom de Google est en fait une modification du nom de «googol» qui en anglais signifie : 10^{100}.
Google est l'une des premières entreprises américaines et même mondiales. C'est l'une des plus imposantes entreprises du marché d'internet. En effet, en 2010, Google possédait un parc de plus d'un million de serveurs, ce qui en fait le parc de serveurs le plus important du monde (2 % du nombre total de machines), avec des machines réparties sur plus d'une trentaine de sites.
Pour de nombreuses personnes, Google est le symbole du monde des services gratuits, performants et sans limites.
Citons un certain nombre d'entreprises ou d'applications développées par google : google maps, google toolba, google sketchUp, Picasa, google desktop, Blogger, google reader, YouTube, Androïd, Chrome, google Analytics, google Earth, Gmail, google chrome OS, Panoramio, igoogle, google livres, streetview, ….
Cependant, la situation de quasi monopole et les questions de vie privée inquiètent de plus en plus, puisque cette société se sert des informations qu'elle récolte pour les revendre ensuite pour de la publicité.
Les revenus Google :

- Google vend des mots clés aux enchères. Si une personne fait une recherche avec ce mot, les liens des sites de ceux qui ont participé aux enchères s'inscrivent dans la partie des liens commerciaux. Chaque fois qu'une personne sélectionne un de ces liens, la société concernée doit verser une certaine somme à Google.
- Lors d'une recherche d'un site Web sur le moteur de Google, les pages des résultats comportent également des annonces publicitaires, annonces vendues par google, sous le terme "liens sponsorisés", et choisies en fonction des mots clés tapés.

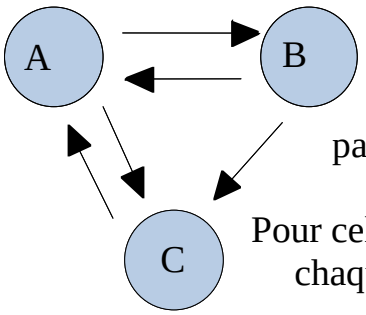
Google utilise, entre autres, un système de classement nommé PageRank pour son moteur de recherche. Ce système est basé sur un algorithme mathématique de popularité. Il a été un succès immédiat et a éclipsé des moteurs de recherche comme Altavista ou Yahoo.

Activité élèves : Partie II :

Initiation aux probabilités par les pourcentages
Comment fonctionne Google.fr ?

> *L'élément fondamental de Google.fr est PageRank, un système de classement des pages Web.*
> *Le principe de PageRank est simple : tout lien pointant de la page A à la page B donne de la popularité à la page B.*
> *L'indice de popularité d'une page est d'autant plus grand qu'elle a un grand nombre de pages populaires ayant un lien vers elle.*
> *Toutefois, Google ne limite pas son évaluation au nombre de liens reçus par la page.*

Un premier exemple :



Supposons qu'il n'y ait que 3 pages sur internet : les pages A, B et C.

Supposons également le schéma (voir ci-contre) de ces pages avec les liens qui les relient : le site Google.fr va calculer la popularité de chaque page.

Pour cela, au départ, attribuons par exemple 10 points de popularité à chaque page, c'est-à-dire que nous supposons que 10 personnes consultent chacune de ces pages.

Chaque page va répartir ses points <u>équitablement</u> entre toutes les pages vers lesquelles elle envoie un lien, autrement dit les 10 internautes qui étaient sur la page vont se répartir équitablement vers les pages sur lesquelles elle envoie un lien.

Nous allons voir sur l'exemple ci-dessus, grâce au tableau suivant, comment vont se répartir ces points, au bout d'un grand nombre de navigations sur le net.

1e étape : départ

A a reçu : 10	B a reçu : 10	C a reçu : 10	Nombre de points
donne	donne	donne	Envoi des points selon les liens
….. à B ; …. à C	…. à A ;	↓	

À la fin de cette étape, il y a …. personnes sur la page A, …. personnes sur la page B, …. personnes sur la page C. Les points de popularité de ces trois pages sont donc, pour la deuxième étape : A …, B …, C …..

À ce stade de la navigation, les pages sont classées dans l'ordre : …., ….., …. et la somme de tous les points de popularité est …. (c'est-à-dire qu'il y a toujours les 30 internautes du départ sur la toile !)

2e étape : On note les points de popularité obtenus à l'issue de la première étape et on recommence le partage des points selon le même principe (chaque page répartit ses points <u>équitablement</u> entre toutes les pages vers lesquelles elle envoie un lien) :

A a reçu : (la moitié des points de B et tous les points de C)	B a reçu : (la moitié des points de A)	C a reçu : (la moitié des points de A et la moitié des points de B)	Nombre de points
donne	donne	donne	Envoi des points selon les liens
			

3e étape : on compte les points obtenus, on note les résultats dans le tableau ci-dessous. À ce stade de la navigation, les pages sont classées dans l'ordre : …, …., … On peut encore vérifier le total de tous les points de popularité : ……
On recommence le partage des points, toujours selon le même principe :

A a reçu : (la moitié des points de B et tous les points de C)	B a reçu : (la moitié des points de A)	C a reçu : (la moitié des points de A et la moitié des points de B)	Nombre de points
donne	donne	donne	Envoi des points selon les liens
			

4e étape : on compte les points obtenus, on note les résultats dans le tableau ci-dessous. À ce stade de la navigation, les pages sont classées dans l'ordre : ..., …, ….
On vérifie le total de tous les points de popularité : ……
On recommence le partage des points, toujours selon le même principe :

A a reçu : (la moitié des points de B et tous les points de C)	B a reçu : (la moitié des points de A)	C a reçu : (la moitié des points de A et la moitié des points de B)	Nombre de points
donne	donne	donne	Envoi des points selon les liens
			

5e étape : on compte les points obtenus, on note les résultats dans le tableau ci-dessous. À ce stade de la navigation, les pages sont classées dans l'ordre : …, …., …
On vérifie le total de tous les points de popularité : ……
On recommence le partage des points, toujours selon le même principe :

A a reçu : (la moitié des points de B et tous les points de C)	B a reçu : (la moitié des points de A)	C a reçu : (la moitié des points de A et la moitié des points de B)	Nombre de points
donne	donne	donne	Envoi des points selon les liens
↙ ↘	↙ ↘	↓	

À ce stade de la navigation, les pages sont classées dans l'ordre : …., ….., ….
On vérifie le total de tous les points de popularité : ……

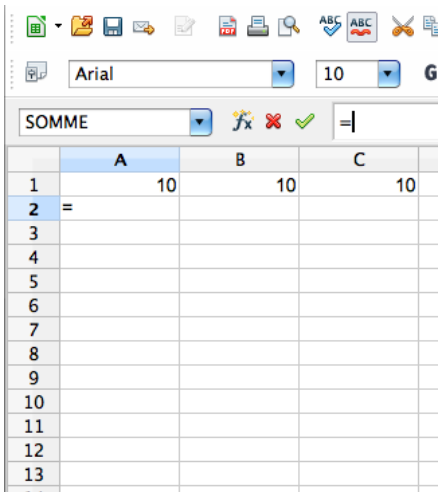
Pour confirmer ce classement nous allons utiliser le tableur mais il faut établir des formules pour lui permettre de calculer ce que reçoit chaque page.

À chaque fois que l'on compte le nombre de points,

A a reçu la moitié des points de B et tous les points de C	B a reçu la moitié des points de A:	C a reçu la moitié des points de A et la moitié des points de B
A =	B =	C =

Mettre ces formules dans le tableur :

Faire vérifier par le professeur



Après un grand nombre d'étapes, les points de popularité ont l'air de se stabiliser et un classement « définitif » se dessine.
Le tableur donne le classement suivant : …, …., ….

Pages		A	B	C
Conditions initiales	au Tableur			
	en Pourcentages			
	en Fréquences			
Résultats après stabilisation	au Tableur			
	en Pourcentages			
	en Fréquences			

Nous allons répartir les points de popularité différemment, et observer les résultats : Que se passe-t-il si on donne tous les points de popularité à A ?

Le tableur donne le classement suivant : …, …, …

Donner tous les points à A c'est donner ….. % des points de popularité à A.

Compléter le tableau suivant :

Faire vérifier par le professeur

Pages		A	B	C
Conditions initiales	au Tableur			
	en Pourcentages			
	en Fréquences			
Résultats après stabilisation	au Tableur			
	en Pourcentages			
	en Fréquences			

Notons qu'il est intéressant de faire remarquer aux élèves qu'en changeant la répartition initiale des points (en mettant par exemple 0 à A, 1 à B, 0 à C, ou en attribuant 0 à A, 0 à B, 1 à C), la stabilisation des points se fait toujours de la même manière : 0,444 ; 0, 2222 ; 0,3333 et que cette répartition est la même en fréquence si on attribue 30 à A, 0 à B, 0 à C, ou encore 30 à A, 10 à B, 45 à C. (voir à ce sujet, le complément d'information donné en annexe par l'éditeur). Cette manipulation se fait

aisément avec le tableur, soit par le professeur au vidéo projecteur, soit par chaque groupe d'élèves. Proposons alors la définition suivante :

Après n étapes, la fréquence de popularité d'une page s'appelle la probabilité de présence d'un internaute sur cette page.

La stabilisation de ces fréquences après un grand nombre de navigations sur le net, permet ainsi de dire que :
La probabilité de présence d'un internaute sur la page A est de :
La probabilité de présence d'un internaute sur la page B est de :
La probabilité de présence d'un internaute sur la page C est de :

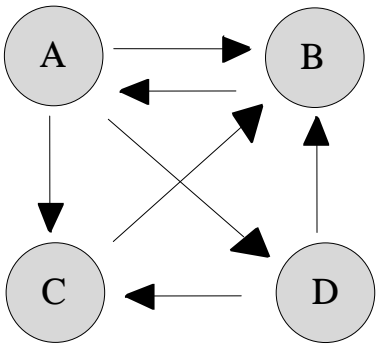
C'est avec ces probabilités que Google donne le classement des pages sur son site de recherche.

Résultats au tableur pour la première configuration proposée aux élèves :

	A	B	C	total		A	B	C	total
étape 0	10	10	10	30		1	0	0	1
étape 1	15	5	10	30		0	0,5	0,5	1
étape 2	12,5	7,5	10	30		0,75	0	0,25	1
étape 3	13,75	6,25	10	30		0,25	0,375	0,375	1
étape 4	13,125	6,875	10	30		0,563	0,125	0,313	1
étape 5	13,438	6,563	10	30		0,375	0,281	0,344	1
étape 6	13,281	6,719	10	30		0,484	0,188	0,328	1
étape 7	13,359	6,641	10	30		0,422	0,242	0,336	1
étape 8	13,320	6,680	10	30		0,457	0,211	0,332	1
étape 9	13,340	6,660	10	30		0,438	0,229	0,334	1
étape 10	13,330	6,670	10	30		0,448	0,219	0,333	1
étape 11	13,335	6,665	10	30		0,442	0,224	0,333	1
étape 12	13,333	6,667	10	30		0,446	0,221	0,333	1
étape 13	13,334	6,666	10	30		0,444	0,223	0,333	1
étape 14	13,333	6,667	10	30		0,445	0,222	0,333	1
étape 15	13,333	6,667	10	30		0,444	0,222	0,333	1
étape 16	13,333	6,667	10	30		0,445	0,222	0,333	1
étape 17	13,333	6,667	10	30		0,444	0,222	0,333	1

Exemple supplémentaire :

En utilisant le tableur, déterminer le classement final de ce schéma à quatre pages :



Classement :	Probabilités :

Réponses : A et B sont ex æquo avec 0,35294118, viennent ensuite C avec 0,17647059, puis D avec 0,11764706. (le lecteur curieux pourra trouver d'autres précisions en annexe.)

Annexe :

Notes et commentaires de l'éditeur :

- ✓ *Pourquoi les probabilités de présence sur une page se stabilisent-elles ? Pourquoi ne dépendent-elles pas de l'état initial ?*

Ces deux résultats viennent d'un même théorème, le fameux Théorème du point fixe : soit S un compact de R^n et soit f une application de S dans S contractante, c'est-à-dire qu'il existe un réel k dans]0 ;1[tel que $\|f(x)-f(y)\| \leq k\|x-y\|$ pour tous x, y de S. Alors :

1. Il existe un unique x_0 de S tel que $f(x_0)=x_0$.
2. Les suites de S définies par $u_{n+1}=f(u_n)$ convergent vers x_0 et ceci quel que soit le choix du terme initial u_0 dans S.

Dans l'algorithme de Google, la dimension de l'espace R^n est le nombre de pages Web. L'application *f* est celle que l'on définit par les formules du tableur. Elle a la propriété d'être linéaire et d'envoyer un vecteur de R^n dont la somme des coordonnées est un certain nombre T sur un vecteur de R^n dont la somme des coordonnées est ce même nombre T. On peut démontrer qu'elle est, **sous de bonnes hypothèses**[1], contractante sur l'ensemble (compact) des vecteurs dont les coordonnées sont positives et de somme T. Les hypothèses du théorème du point fixe sont donc satisfaites.

Attribuer au départ un nombre total de points de popularité T aux pages web revient à choisir un vecteur u_0 dont la somme des coordonnées est T. Les points de popularité des pages après le $n^{ième}$ clic sont les coordonnées du vecteur u_n défini par récurrence par $\boldsymbol{u_{n+1}=f(u_n)}$. Le théorème du point fixe implique que, quelle que soit la répartition initiale des points entre ces différentes pages, la suite (u_n) converge vers l'unique point fixe de f dont la somme des coordonnées est T. Si on calcule les probabilités de présence et non plus les points de popularité, l'algorithme converge vers l'unique point fixe de f dont la somme des coordonnées est 1 (qui, f étant linéaire, est obtenu simplement en divisant par T les coordonnées du point fixe de f dont la somme des coordonnées est T).

- ✓ *Probabilités de présence (ou points de popularité) au bout d'un temps infini de l'exemple à trois pages Web de l'activité-élève, partie II :*

[1] Il faut que la seule valeur propre de module 1 de l'application linéaire f soit 1 et que le sous-espace propre associé soit de dimension 1.

D'après ce qui précède, pour déterminer les probabilités de présence au bout d'un temps infini ou les points de popularité au bout d'un temps infini, il faut trouver les points fixes de l'application f de R^3 dans R^3 définie par les formules destinées au tableur. Notons x_A, x_B, x_C les valeurs des pages A, B, C. Cette application est alors définie par :

$$f(x_A ; x_B ; x_C)=(0{,}5x_B+x_C ; 0{,}5x_A ; 0{,}5x_A+0{,}5x_B).$$

Les points fixes de f sont solutions de l'équation $f(x_A ; x_B ; x_C)=(x_A ; x_B ; x_C)$. Ce sont donc les triplets $(x_A ; x_B ; x_C)$ qui satisfont aux trois équations :

$$\begin{cases} 0{,}5x_B + x_C = x_A \\ 0{,}5x_A = x_B \\ 0{,}5x_A + 0{,}5x_B = x_C \end{cases}$$

c'est-à-dire les triplets de la forme $x_A = 2\alpha$; $x_B = \alpha$; $x_C = 1{,}5\alpha$; avec α réel.

- Probabilités de présence au bout d'un temps infini : on les obtient en cherchant la solution qui vérifie $x_A + x_B + x_C = 1$, soit $\alpha = 2/9$ et la solution est :

  $x_A = 4/9$; $x_B = 2/9$; $x_C = 1/3$.

- Points de popularité : si on répartit 30 points de popularité entre les 3 pages, il faut chercher la solution telle que $x_A + x_B + x_C = 30$, c'est-à-dire prendre $\alpha = 20/3$ et les points de popularité au bout d'un temps infini sont : **$x_A = 40/3$; $x_B = 20/3$; $x_C = 10$.**

De même, nous aurions le système associé à l'exemple supplémentaire de la partie II de l'activité élève :

$$\begin{cases} x_A = x_B \\ x_B = \frac{1}{3}x_A + x_C + \frac{1}{2}x_D \\ x_C = \frac{1}{3}x_A + \frac{1}{2}x_D \\ x_D = \frac{1}{3}x_A \end{cases}$$

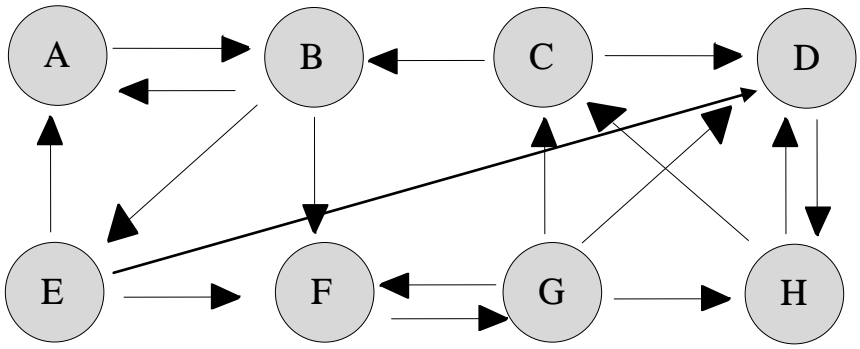
Pour lequel les probabilités de présence sur une page sont : $x_A = x_B = \frac{6}{17}$; $x_C = \frac{3}{17}$; $x_D = \frac{2}{17}$.

Pour aller plus loin, un point de vue matriciel : la matrice de l'application linéaire f modélisant le réseau est :

$$M = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix}$$

L'égalité $u_{n+1}= f(u_n)$ se traduit matriciellement par $X_{n+1} = M\ X_n$ où X_n est le vecteur $(x_A ; x_B ; x_C ; x_D)$. Soit X_0 le vecteur contenant les valeurs initiales que l'on se fixe au hasard. On a alors $X_n = M^n\ X_0$. Pour n assez grand, les coefficients de M^n se stabilisent et le point fixe sera approché par M^nX_0.

Autre exemple, réseau à 8 pages :



La matrice correspondant à ce réseau est :

$$M = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 \\ 1 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{3} & 0 & \frac{1}{4} & \frac{1}{2} \\ 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{4} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \frac{1}{4} & 0 \end{pmatrix}$$

Les probabilités de présence sur chaque page sont : $\frac{2}{35}$; $\frac{9}{70}$; $\frac{1}{7}$; $\frac{8}{35}$; $\frac{3}{70}$; $\frac{8}{105}$; $\frac{8}{105}$; $\frac{26}{105}$.

Références :

1- *Mathématiques et Technologie* de C. Rousseau et Y. Saint-Aubin (livre édité par Springer-Verlag) disponible à l'IREM.

2- *Comment fonctionne Google* de M. Eiserman (archive de Wikipédia sur PageRank)