1 Introduction

Ces dernières décénies ont été connues pour le bond extraordianaire que les réseaux ont eu, dans le monde scientifique : tels que les réseaux de proteines, dans le monde industriel grâce aux réseaux de neurones où tout simplement dans la vie de chacun d'entre nous grâce aux réseaux sociaux ou le réseau du Web.

Ici, dans ce mémoire, nous nous consacrerons à l'un de ces réseaux, à savoir le Web, immense réseau modélisé précédement par les chercheurs comme étant un graphe dont les noeuds représentes les pages web et les arcs, les liens hypertextes les reliant. Immense réseau mais surtout réseau en pérpétuelle expension ce qui peut rendre la recherhe d'information par les moteurs de recherche un peu difficile.

Ce modèste travail consistera à étudier le Web dans le but de le modéliser par des graphes probabilistes tout en gardant ses caractéristiques existantes déjà à savoir la distribution de degrés en loi de puissance, une faible distance moyenne et un fort coêfficient de clustring et ce afin de faciliter le travail des moteurs de recherche (trouver le document voulu en un minimum de temps). Dans ce modèste résumé, une présentation sera faite sur ce qu'est concrétement le Web, du point de vu structural, ou topologique afin de bien modéliser, puis nous présenterons quelques modèles de réseaux pour pouvoir comparer, enfin nous décrirons notre modèle.

2 Qu'est ce que le Web?

Abréviation de World Wide Web ou grande toile d'araignée mondiale, le Web est un outil permettant de trouver des documents ou un ensemble d'informations rendus public et de consulter ainsi toute sorte de pages : actualités, recherche, loisirs, art ...

Il est caractérisé par :

1. Un langage de programmation, connu sous le nom de html utilisé pour écrire les pages Web et d'insérer des hyperliens dans du texte.

2. Un protocole http : protocole de transport et de diffusion de tout type de fichier : texte, images, sons, vidéos...afin qu'ils soient interprétés du côté client par des navigateurs Web les Browsers.

3. Sa fluctuation permanante:

- 26millions de pages en 1998 (d'après google).
- 1milliard de pages en 2000.
- 8 168 684 336 de pages en 2005.
- 1000 milliards de pages en 2008.

La localisation d'un document sur ce dernier devient une tâche assez complexe. On parle alors de visibilité et d'accessibilité du Web[2, 3, 4, 5] où Le Web visible est tout simplement tous les sites dont le contenu peut être aspiré par un robot comme par exemple celui des grands moteurs comme Google ou Yahoo quant au Web invisible représentetoutes les sources dont l'accès est contrôlé telle que :

- Les bases de données.
- Les bibliothèques en ligne payantes.
- L'indexation.
- Les documents dynamiques.
- Le référencement.

et l'accessibilité comme son nom l'indique, représente le Web auquel on a facilement acces ou pas. Les causes d'inaccessiblité du Web sont principalement les même causes du Web invisible.

Dû à cette visibilité et accessibilité plus ou moins facile du Web, notre travail consistera à modéliser le Web en partie afin de le rendre plus accessible qu'il ne l'est dejà.

3 Structure du Web

L'étude structurale du Web est une étape très importante pour la modélisation de notre modèle.

Nous en definissons les deux principales à savoir la structure microscopique et la microscopique.

3.1 La vision macroscopique

Cette structure a été établie en 1998, définie comme étant un noeud papillon composé de 4 parties :

La partie centrale (noyau fortement connexe):
Appelée aussi le cœur du Web, cette partie représente 27, 2% de la totalité du graphe soit moins d'un tiers des pages Web. La navigation est aisée, on peut aller de n'importe quelle page à n'importe quelle autre. Son diamètre est d'au moins 28. De plus la longueur moyenne d'un plus court chemin est comprise entre 16 et 20.

La partie appelée " Origine ou IN ":
Soit 21, 5% du total (environ 1/5). Les pages la constituant peuvent être reliées entre elles. L'accès au cœur du Web est faisable mais l'inverse n'est pas possible. Ce sont sans doute de nouveau sites que les internautes n'ont pas encore découverts.

La partie " Extrémités ou OUT" (pages de destination):
Représente aussi 21,5%. Accessible depuis le noyau mais aucun retour n'est possible. On peut trouver des sites commerciaux vers lesquels pointent de nombreux liens mais qui, eux, n'en propose pas, ou seulement en interne.

Des " Branches ou des TENDRILS ":
Il n' y a aucun échange avec le noyau mais qui sont accessible à partir de la zone " Origine " et accèdent à leur tour à la zone " Extrémité ".

Les pages déconnectées:
Ce sont des pages qui n'ont aucun lien externe, elles peuvent être des pages personnelles. Cette partie représente 10

3.1.1 Caractéristiques de la structure macroscopique

Les résultats obtenus sont les suivants :

- Le diamètre du graphe en entier est de 500.
- La distance moyenne entre une page u de l'ensemble Origine et v de l'ensemble Extrémité en passant par le noyau est proche de 900.
- La probabilité qu'un chemin existe entre deux pages quelconques est de 24%.

Pour cette étude, les expérimentations ont été effectuées à partir de données provenant du moteur de recherche AltaVista (203 549 046 pages Web et 1, 5 milliard de liens hypertextes).

3.2 La vision microscopique

Cette étude se focalise sur les particularités locales et tout particulièrement sur les communautés qu'il faut savoir définir et détecter.
Une première définition d'une communauté sera décrite comme étant un

couple d'ensemble de pages Web tels que toutes les pages du premier ensemble (appelées " fans ") pointent vers toutes les pages du second (stars) qui ne pointent pas les unes sur les autres. Notons que cette communauté est centrée autour d'un sujet de prédilection.
Exemple :
Des passionnés de voyage pointeront vers des pages d'agences de voyage, mais qui, elles ne pointent les une sur les autres car elles sont concurrentes.
Si nous comparons cette définition avec l'algorithme HITS (Jon Kleinberg), nous verrons que les fans représentent les pivots et les stars, les pages qui font autorité. Une autre définition décrit la communauté comme une collection de pages Web qui possèdent plus de liens hypertextes entre les pages de la collection qu'avec les pages externes.

4 Caractéristiques du Web

Le Web se caractérise par principalement trois propriétés:

1. **La distribution de degrès en loi de puissance**
   Cette distribution représente la probabilité qu'un sommet ait un degré d, est d'un ordre proportionnel à $1/d^x$, avec x>1. Ce qui signifie que les sommets avec un petit degré sont beaucoup plus nombreux que ceux dont le degré est élevé.
   La figure(4) montre :

   - À gauche : la distribution des degrés entrants, prise en deux fois (mai 99, octobre 99), a une certaine homogénéité avec la loi de puissance dont l'exposant est égale à 2,09.

   - Au centre : concerne les degrés sortants dont les distributions (celles demai et octobre) dévie quelque peu de la loi de puissance (exposant =2,72). Mais en somme, les deux cas, les distributions des degrés entrants et sortants et la loi de puissance sont similaires.

   - À droite : la loi de puissance relative à la distribution des degrés entrants quant à elle, est comparée à la loi de Zipf(loi qui fait référence à la taille et à l'apparition d'un événement (un mot en l'occurrence dans l'expérience de George Kingsley) par rapport à son rang).

2. **La distance moyenne**
   soit G = (V, E) un graphe quelconque.
   Notons que :
   En considerant G comme etant le graphe du Web, certaines recherches

ont determine cette distance moyenne de l'ordre de 19, c'est à dire qu'à partir de n'importe quelle page Web, nous pouvons atteindre n'importe quelle autre en suivant en moyenne 19 liens :
definie par R. Albert, H. Jeong et A.L. Barabasi dans l'une de leurs recherches faites en 1999, comme étant le plus court chemin entre deux documents ou comme le plus petit nombre de liens d'URLs devant etre suivi afin de surfer d¡¯un document a un autre, cette distance moyenne du graphe du Web a travers tous les sommets, notée <d>, a été trouvee égale à : <d>= 0;35+2, 06log(N.
Les résultats ont montré aussi que pour un N donné, d suit la loi de Gauss, ainsi <d> peut etre interpretée comme étant le diametre du Web.
Pour $N = 8 \times 10^8 \Rightarrow distance_{Web} = 18,59$.
Ce qui signifie que deux documents pris aléatoirement sur le Web sont à une distance moyenne de 19 cliques l'un de l'autre.
Cette petite valeur de <d>, indique qu'un agent intelligent pouvant interpréter les liens et suivre uniquement ceux qui sont pertinents, peut trouver rapidement l'information désirée en surfant sur le Web, ce qui n'est pas le cas pour un robot qui localise l'information basée sur le matching strings.
Ces mêmes recherches ont trouvé qu'un tel robot visant à identifier un document à distance <d> a besoin de chercher M <d> implique

$$0,53 * N^{0,92}$$

documents , qui avec

$$N = 8 * 10^8$$

,suit

$$M = 8 * 10^7$$

ou 10% du Web.
A fin de vérifier la validité des prédictions, d fut déterminé à partir de document pris du domaine nd.edu.
La distance mesurée

$$< d_{nd.edu} >= 11.2$$

rejoint celle prédit

$$< d_{3\times10^5} >= 11.6$$

obtenue du modèle.

3. **Le coefficient de clusterisation**
   Une mesure introduite en 1998 par Ducan J.Watts et Steven Strogatz afin de voir si un graphe est un réseau petit monde, le coefficient de clusterisation d'un sommet dans un graphe quantifie combien celui ci

et ses voisins sont au cours d'une clique.
Plus précisément, le coefficient de clusterisation Ci pour un sommet

$$v_i$$

est la proportion de liens existant entre ses voisins divisée par le nombre de liens qui peut exister entre eux.
Ainsi, le coefficient de clusterisation pour un graphe orienté est défini comme suit :

$$C_i = \frac{|\ \{e_{jk}\}|}{k_i(k_i - 1}; v_j, v_k \in N_i, e_{jk} \in E$$

où

$$k_i$$

est le nombre de degrés du sommet i;
pour celui du graphe non orienté, il est donné par :

$$C_i = \frac{|\ \{2 * e_{jk}\}|}{k_i(k_i - 1}; v_j, v_k \in N_i, e_{jk} \in E.....(1)$$

soit :

$$\lambda_G(v)$$

: le nombre de triangle sur v pour un graphe non orienté. C'est aussi le nombre de sous graphes de G ayant 3 arêtes et 3 sommets où v est l'un d'entre eux.

$$\tau_G(v)$$

: le nombre de triples sur v ou le nombre de sous graphes ayant 2 arêtes et 3 sommets, l'un d'eux est v, tel que v soit incident aux deux arêtes.
De ces deux quantités, le coefficient de clusterisation peut être défini par :

$$C_i = \frac{\lambda}{\tau}.....(2)$$

$$\tau_G(v) = C(k_i, 2) = \frac{1}{2}k_i(k_i - 1)$$

Ce coefficient représente aussi la probabilité pour que deux voisins d'un même sommet soient eux même voisins sachant que :
Le coefficient de clusterisation d'un sommet v de G est la probabilité étant donné deux voisins v' et v" qu'ils soient voisins entre eux, le coefficient de clusterisation du graphe G est la moyenne du coefficient de

clusterisation de tous ses sommets (défini par Watts et Strogatz) :

$$\overline{C} = \frac{1}{n} \sum C_i$$

Remarque
Il existe ainsi deux définitions pour le coefficient de clusterisation:

(a) une définition globale:

$$C_1 = \frac{\sum_{u \in V} nombredetrianglesdontuestunsommet}{\sum_{u \in V} nombredetripletsdontuestunsommet}$$

(b) une définition basée sur la moyenne de coefficients locaux:

$$C_2 = \frac{1}{n} \frac{\sum_{u \in V} nombredetrianglesdontuestunsommet}{\sum_{u \in V} nombredetripletsdontuestunsommet}$$

La figure 6 représente un exemple sur le coefficient de clusterisation sur un graphe non orienté sur le sommet vert.
Les arêtes en pointillé sont pour les arêtes potentiellement inutilisables.
Les arêtes bleues quant à elles relient les voisins du sommet vert.

4. **Small World**
   Combinaison de la distance moyenne et du coefficient de clusterisation, un réseau petit monde ou small world est défini aussi comme étant un graphe où la plupart des noeuds ne sont pas voisins entre eux, mais la plupart d'entre eux peuvent être atteints par n'importe quel sommet par un petit nombre de clique.
   Ainsi, un graphe est considéré comme small world si son coefficient de clusterisation moyen C est significativement plus grand que celui dans un graphe aléatoire construit sur le même ensemble de sommet et si le graphe a une faible distance moyenne.

5 Modèles de Réseaux

Dans cette partie, nous présentons les deux modèles principaux utilisés pour la construction de notre modèles:

5.1 Le modèle de Watts et Strogatz

Le modèle de Watts et Strogatz joue un rôle important dans la modélisation des topologies réalistes.
Il possède d'intéressantes propriétés :

un fort coefficient de clusterisation ainsi qu'une distance moyenne faible.
Ce modèle convient donc à la propriété small world.
Il est défini comme suit :
Il interpole entre un treillis régulier et un réseau aléatoire.
Le modèle prend un paramètre singulier p et produit un réseau selon l'algorithme suivant:

Commencer par ordonner Disposer un anneau en treillis dans lequel il y a N nœuds, chacun connecté avec ses k voisins $(k/2)$ de chaque coté. Afin d'avoir une faible densité, mais bien relié au réseau de tous les temps, nous considérerons $N \gg K \gg ln(N) \gg 1$

Randomiser Des connexions aléatoires longue distance sont ajoutées grâce à la procédure suivante: 1.Visiter chaque nœud.
2.À un nœud donné, visiter chaque lien.
3.Recabler cette connexion en la déplaçant vers un autre nœud sélectionné aléatoirement (probabilité uniforme tout en évitant l'auto connexion et la duplication de liens) avec une probabilité p.

5.2 Les réseaux Scale free

Albert et Barabasi démontrèrent en 1998 qu'indépendamment de la nature du système et de l'identité de ses composants , la probabilité P(k) qu'un sommet dans un réseau soit connecté à k autres sommets décroît comme une loi de puissance, suivant $P(k) \sim k^{-\gamma}$, un phénoèmen nommé scale invariance ou scale free
La caractéristique générique de cette observation fut supportée par 4 exemple " real world ":

1. Le graphe des acteurs.
2. Le graphe de la grille de puissance.
3. Le graphe des publications scientifiques.
4. Le graphe du Web.

Ces réseaux Scale free se caractérisent par :

- Une distribution de degrés suivant une loi de puissance $P(k) \sim k^{-\gamma}$ avec $2 < \gamma < 3$.
- Une distance moyenne proportionnelle à log(n)/log(log(n)).
- Un coefficient de clustring constant.

6 La Modèlisation

Dû à la grande taille du Web, sa représentation est généralement sous forme graphique où G=(X,E) où X sera l'ensemble des pages HTML et E l'ensembles des liens reliants ces pages.
Le travail demandé étant de modéliser le Web par des graphes probabilistes, ces derniers devront satisfaire les caracteristiques dejà existantes dans le Web à savoir :

- La distribution de degrés en loi de puissance.
- Une faible distance moyenne.
- Un fort coefficient de clustring.

6.1 Comment obtenir la distribution de degrés en loi de puissance ?

Comme nous l'avons vu précédement, le principe de base de la distribution de degrés en loi de puissance repose sur le fait qu'il y ait peu de pages ayant un fort degré et beaucoup de pages avec un degré faible.
Pour cela, nous avons choisi de partager le Web en communautés. Ainsi dans notre modèle , l'ensemble X ne sera plus l'ensembles des pages HTML mais l'ensemble des communautés et E deviendra l'ensemble des liens reliants ces communautés.
Nous auront ainsi peu de pages avec beaucoup de liens, ces pages sont bien evidemment les pivots des commuanutés et beaucoup de pages ayant moins de liens que les pivots.

6.2 Comment obtenir une faible distance moyenne et un fort coefficient de clustring ?

Sachant que la combinaison de ses deux propriétés forme un small world, nous avons choisi de prendre comme support pour notre modèle , le graphe probabiliste de Watts et Stroagtz qui nous le rappellons est un treillis de sommets reliés entre eux de façon à avoir un fort coefficient de clustring et une faible distance moyenne, selon une probabilité comprise entre 0 et 1.
La figure 10 nous donne un aperçu du modèle.

7 Conclusion

La particularité de notre modèle est qu'il réduit d'une manière surprenant (de 18 à 3) la distance moyenne et auguemente le coefficient de clustering

tout en gardant la distribution de degrès telle quelle.

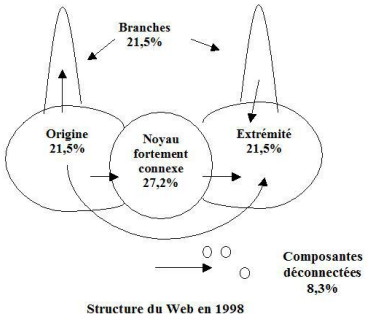

Figure 1: Vision macroscopique du Web

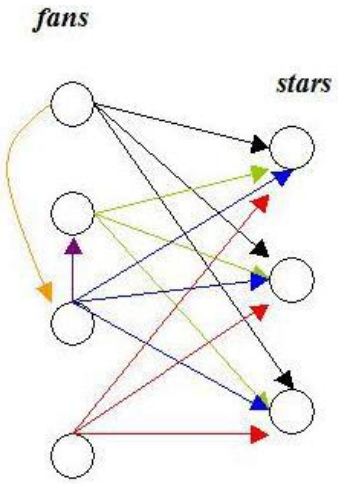

Figure 2: Vision microscopique du Web

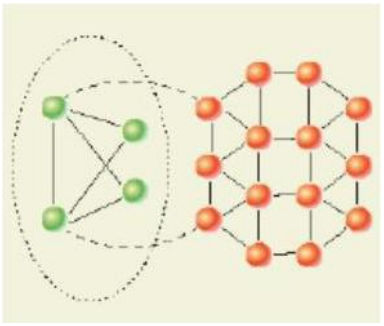

Figure 3: Autre ision microscopique du Web

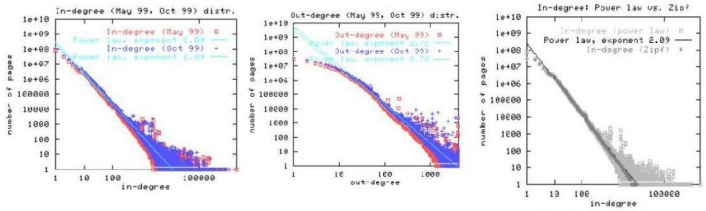

Figure 4: à gauche :Distribution des degrés entrants. centre : Distribution des degrés sortants. à droite : Loi de Zipf

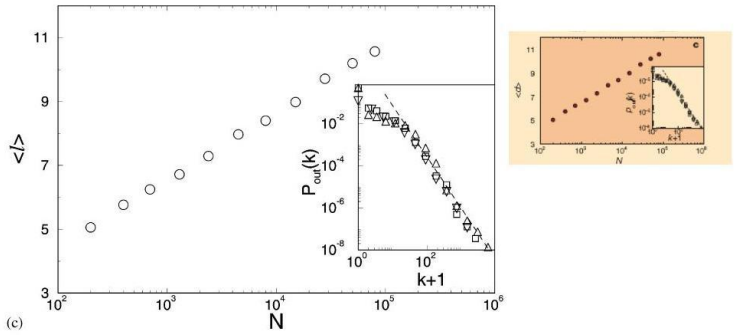

Figure 5: Distance moyenne entre deux documents comme fonction de la taille du système comme prédit par le modèle

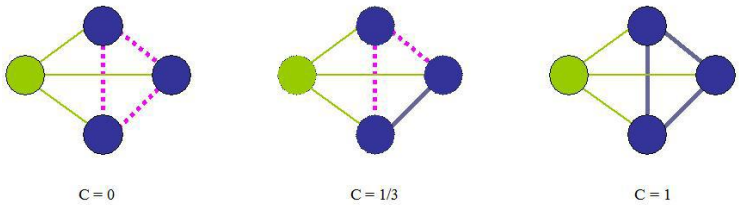

Figure 6: Représentation du coefficient de clustering.

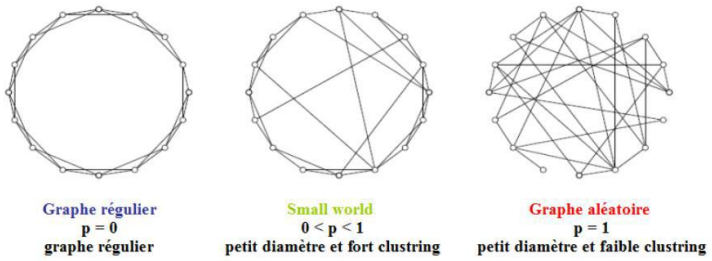

Figure 7: Différents types de graphes selon le changement de probabilité.

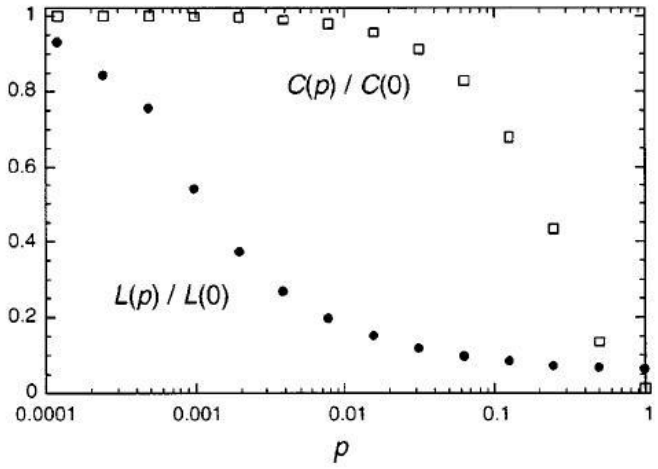

Figure 8: Courbes de la distance moyenne et du coefficient de clustring

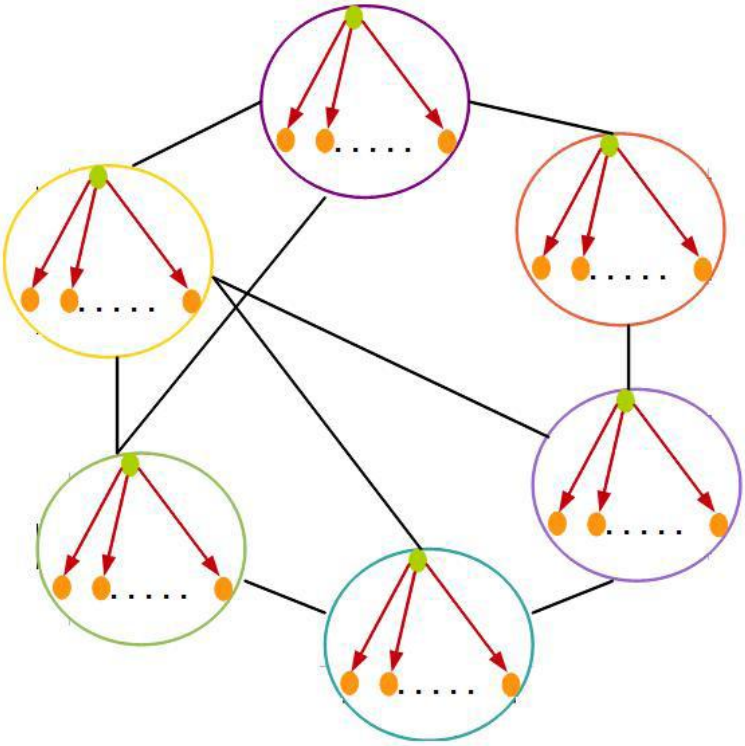

Figure 9: Modèle proposé pour la modélisation du Web respectant les propriétés du Web initial.