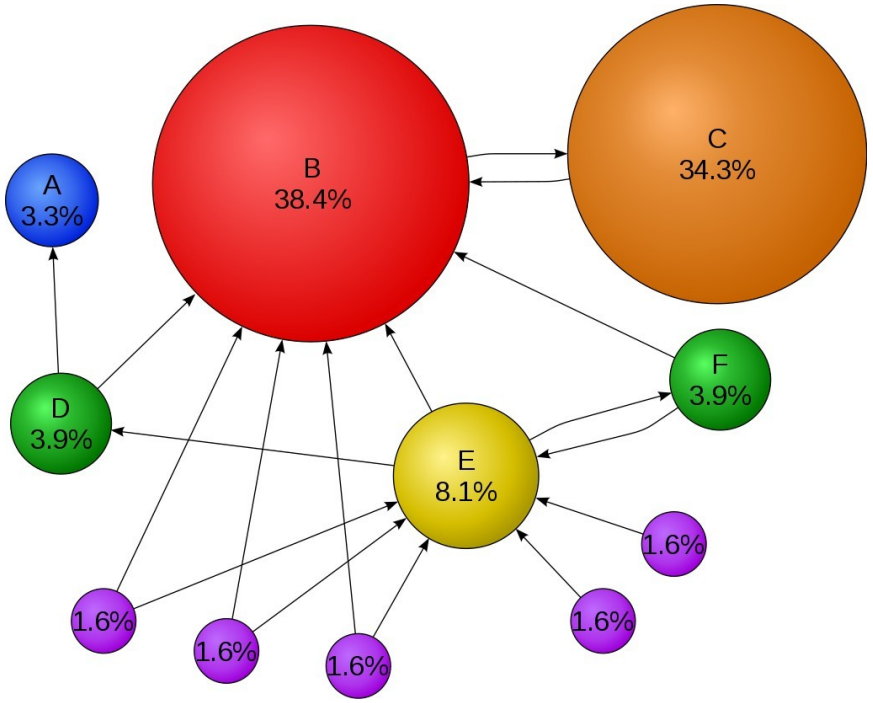
Comprendre le PageRank

(sources : wikipedia.org, académie de Franche-Comté)

La méthode consiste à évaluer la notoriété d'une page web. Elle a été inventée par Larry Page et Sergey Brin, brevet déposé en Janvier 1997 et enregistré le 9 Janvier 1998 (Method for Node Ranking in a Linked Database). Cette méthode leur a permis de lancer leur moteur de recherche Google deux mois plus tôt. En quelques années le moteur est devenu le plus célèbre du web.



Le principe revient à modéliser le web comme un graphe dont les pages sont les nœuds (ronds) et les hyperliens les arrêtes (flêches). A chaque page est associé un nombre positif entre 0 et 1, appelé score de la page (en anglais "PageRank"). Le score doit rendre compte des deux règles suivantes :

> **R1 :** le score attribué a une page doit être d'autant plus élevé que celle-ci est référencée dans une page faisant autorité (dont le score élevé).
> **R2 :** le score attribué a une page doit être d'autant moins élevé que celle-ci est référencée dans une page contenant un grand nombre de références.

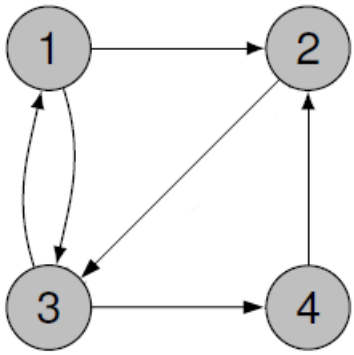
Pour évaluer le score des pages, on imagine un internaute qui se déplace aléatoirement de page en page. Ainsi le résultat ne dépendra que des liens que reçoit et émet le site. En mathématiques on dit que l'internaute est un marcheur aléatoire qui parcoure le graphe. Le procédé suivant peut être appliqué :

1. Choisir un nœud de départ.
2. Marquer le nœud comme visité une fois.
3. Déterminer au hasard (lancé de dé, ...) une arrête à suivre.
4. Se déplacer au nouveau nœud.
5. Recommencer la procédure à l'étape 2.

Le PageRank des nœuds correspondra à la fréquence de leur visite par le marcheur aléatoire. Plus la procédure sera répétée, plus l'évaluation sera précise.

Etude d'un premier cas

Le robot d'un moteur de recherche (spider, crawler) a permis d'établir les relations suivantes entre quatre pages web. Elles sont modélisées dans le graphe ci-dessous.



Dans ce graphe, la flèche allant de 1 vers 2 signifie que la page 1 référence la page 2 et l'absence de flèche de 2 vers 4 signifie que la page 2 ne référence pas la page 4.

Appliquer la méthode du marcheur aléatoire en complétant le tableau ci-dessous. On travaillera en binôme et on effectuera 20 visites.

	Noeud 1	**Noeud 2**	**Noeud 3**	**Noeud 4**
Nombre de visites (sur un total de 20)				
PageRank				

Comparer vos résultats avec ceux des autres binômes de la classe. Que peut-on en conclure ?
.....
.....
.....
.....

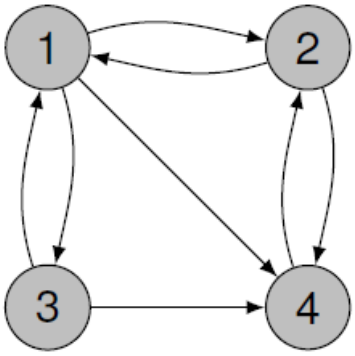
Rassembler les résultats de toute la classe et compléter le nouveau tableau ci-dessous.

	Noeud 1	**Noeud 2**	**Noeud 3**	**Noeud 4**
Nombre de visites (sur un total de …...)				
PageRank				

Reporter le PageRank en % sur le graphe.

Etude d'un deuxième cas

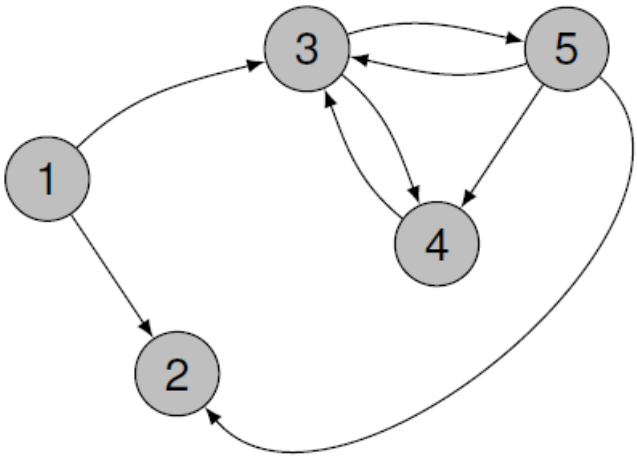
Faire de même avec le cas ci-dessous.



	Noeud 1	**Noeud 2**	**Noeud 3**	**Noeud 4**
Nombre de visites (sur un total de …...)				
PageRank				

Un premier problème

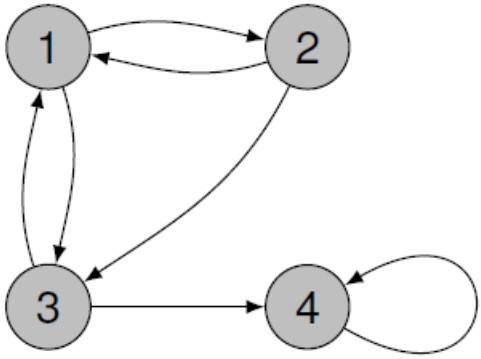
Deux nœuds posent problème dans le graphe ci-dessous.



Le noeud …. :.....
.....
.....
.....

Le noeud …. :.....
.....
.....
.....

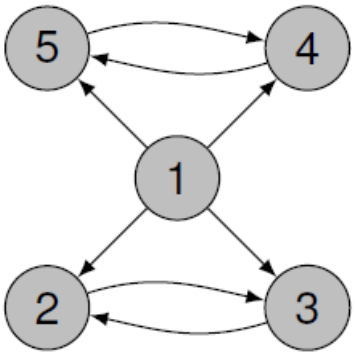
Un deuxième problème : le puit



Quel problème cet exemple soulève-t-il ?

.....
.....
.....
.....

Un troisième problème : la poche



Quel problème cet exemple soulève-t-il ?

.....
.....
.....
.....

La solution de Google :

Si un nœud ne comporte aucun lien vers l'extérieur, le marcheur saute aléatoirement vers un des autres nœuds. Cela simule un internaute qui se trouverait dans ce cas de figure. Bloqué sur un site, il en changerait tout simplement.

De plus, même s'il n'est pas bloqué, à chaque nœud le marcheur a une probabilité de 0.15 (15%) de sauter aléatoirement vers un des autres nœuds. Cela simule l'ennuie de l'internaute à suivre toujours les liens proposés par les sites.

On peut démontrer que ces propositions sont valables mathématiquement en théorie des graphes orientés et donne les bons résultats limites.

Recalculer les PageRanks des graphes précédents avec cette méthode et indiquer les valeurs obtenues dans les nœuds.