



BIG DATA : UNE NOUVELLE FORME D'INTELLIGENCE COLLECTIVE

Mini-série de Billets #4 : Big Data – Sur les épaules de géants

Billet #4B : Le graphe : une pierre angulaire du Big Data

L'évolution qui part des moteurs de recherche et se poursuit jusqu'au Big Data constitue le fil rouge de cette série *Le Big Data : une nouvelle forme d'intelligence collective*. Fil rouge tressé d'un brin incontournable : l'apport essentiel amené par les mathématiques via la théorie des graphes et la topologie. Ainsi que leur rencontre avec l'informatique, matérialisée par le moteur de recherche Google et l'algorithme de classement PageRang.

Ce sujet nous occupera tout au long de ce Billet #4B appréhendé comme une promenade, en clin d'œil à celui qui nous servira d'éclaireur : Leonhard Euler, mathématicien suisse, dont les lumières sous-tendent notre monde numérique. En résolvant au XVIII^e siècle le casse-tête de la promenade des *Sept ponts de Königsberg*, il a posé les premiers jalons d'une approche inédite de la géométrie qui nous permet aujourd'hui de trouver en quelques clics une multitude d'informations.

Notre marche de ce jour à travers l'espace et le temps, effectuée sur les épaules de géants d'époques différentes se divisera en deux grandes étapes :

- Du croquis au graphe
- Graphe et Big Data

Sur les épaules de géants. Une ascension. Nous monterons. Et au sommet, la vue à 360° vous appartiendra, perspective toute personnelle sur notre monde en pleine mutation.

DU CROQUIS AU GRAPHE

Le Web, un graphe de forme familière.

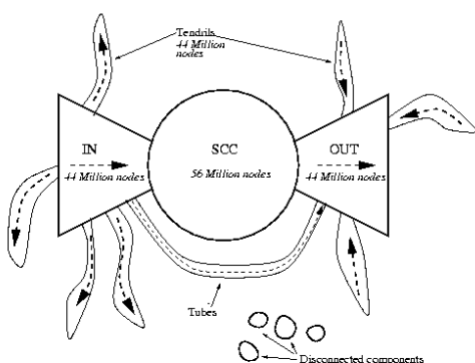
Farfalle, papillons en italien. Cette forme si familière, retrouvée dans nos assiettes, dans nos



penderies, mais surtout dans la nature, sous-tend nombres de nos activités et habitudes : surfer sur les réseaux sociaux, chercher une information sur notre moteur de recherche préféré, par exemple. Et oui, visuellement et schématiquement, c'est à cela que



Graph structure in the web



ressemble le Web, composé de milliards de pages et de sitesⁱ : à un nœud papillon. En dépit de sa complexité, en 2000, des chercheursⁱⁱ ont néanmoins réussi à surmonter la difficulté de le représenter visuellement. La forme mise à jour a révélé « *un graphe en forme de nœud papillon (...) principalement constitué d'un noyau interconnecté et d'autres composantes reliées à ce noyau* »ⁱⁱⁱ. A la question de savoir, pourquoi il est important de connaître la structure générale du Web, les auteurs de *Graph structure in the web* répondaient : « *The study of the web as a graph is not only fascinating in its own right,*

but also yields valuable insight into web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution^{iv}. *We report on experiments on local and global properties of the web graph using two Altavista crawls each with over 200 million pages and 1.5 billion links* »^v.

The study of the web as a graph... Ils ont procédé en étudiant le Web en tant que graphe.

Qu'est-ce, en substance ?

La théorie de graphes et la topologie. En mots.

Selon Napoléon Bonaparte, « *Un bon croquis vaut mieux qu'un grand discours* ». Le premier, à l'instar du graphe, fait appel à la vue, à la visualisation immédiate, à une organisation minimaliste de l'essentiel alors que le second sollicite en premier lieu l'ouïe, trouve sa profondeur dans le détail et la force des mots. Non-immédiat, il s'inscrit inévitablement dans une certaine durée. Alors, un bon croquis vaut-il vraiment mieux qu'un grand discours ? La préférence judicieuse en faveur de l'un ou l'autre mode d'expression relève, me semble-t-il, des circonstances particulières.

Les graphes ont fait l'objet d'une discipline, mathématique, puis informatique : la théorie des graphes. Ils se définissent comme des structures constituées de nœuds (ou sommets, en référence aux polyèdres^{vi}) reliés entre eux par des arêtes (liens, arcs), dont la combinaison constitue une figure, un modèle abstrait, qui ressemble à un réseau. Ils peuvent être de types variés selon leur forme (topologie) et leurs propriétés^{vii}.

Selon leur topologie... Les graphes possèdent une forme, une topologie.

La topologie, constitue également une branche des mathématiques, qui bien que peu connue du grand public, connaît néanmoins une vraie *success story* depuis le XXe siècle^{viii}.

Utilisée aujourd'hui essentiellement de manière indirecte – dans l'étude de l'ADN^{ix}, du fonctionnement du cerveau^x, au croisement des mouvements du corps et de l'art^{xi}, par exemple. Science dédiée à l'étude « *des propriétés géométriques immuables malgré les déformations continues. On y trouve par exemple les surfaces, les nœuds et les liens* »^{xii}, elle « *s'intéresse plus précisément aux espaces topologiques et aux applications qui les lient, dites « continues »*. Elle permet de classer ces espaces »^{xiii}.

A ne pas confondre avec la topographie^{xiv}, « *la topologie est peu connue car elle n'est pas enseignée à l'école. Elle s'intéresse à la notion de formes, mais pas comme en géométrie, où il est question de métriques. Dans la topologie, on imagine qu'on travaille avec de la pâte à modeler pour se figurer quelles sont les propriétés des formes et celles qui seront conservées à la suite des déformations* » expliquait Kathryn Hess, professeure à l'EPFL, dans une interview accordée au journal *Le Temps*^{xv}.

Dénommée à ses premières heures *geometria situs* puis *analysis situs*^{xvi}, la topologie est appelée aussi *géométrie du caoutchouc* parce qu'elle constitue « *une sorte de géométrie de figures tracées sur une surface déformable, comme une feuille de caoutchouc* »^{xvii} afin « *que les droites puissent se plier, rétrécir ou s'étendre, et que les cercles puissent s'écraser pour devenir un triangle ou un carré* »^{xviii}. *Géométrie du caoutchouc* - on se représente bien grâce à cette expression en quoi cette discipline se distingue de la géométrie traditionnelle remplie de longueur, d'angles, de triangles, de ronds et des carrés.

En royaume de topologie, la continuité est souveraine^{xix}. La continuité, aspect fondamental du monde naturel, et caractéristique essentielle des mathématiques^{xx}. De la genèse et du développement du Big Data, aussi.

Avec un peu d'imagination il est possible, dans cet exemple de la déformation continue d'un *mug* en un *donut*, et inversement, de se représenter à l'œuvre les concepts topologiques clés de forme, de déformation et de continuité. La transformation ici statique est disponible en animation sur Wikipédia^{xxi}.



Je vous propose maintenant un voyage dans le temps pour découvrir des géants et le rôle qu'ils ont joué en relation avec la théorie des graphes et de la topologie, dans l'apport des mathématiques à l'amélioration des moteurs de recherche, au Big Data, et au Numérique en général.

Sur les épaules de géants.

Tout comme « *Rome ne s'est pas faite en un jour* », notre société numérique est le fruit d'une évolution, d'une transformation progressive, jalonnée de découvertes, d'innovations, de fulgurances intellectuelles, constitutives d'une évolution scientifique, historique, humaine. Emanations de divers domaines. Elle n'est pas advenue spontanément. A l'instar de l'arbre, notre société numérique a des racines multiples, plongeant plus ou moins profondément dans le sol, l'ancrant plus ou moins solidement dans l'Histoire. Et l'évolution se poursuit.

« *If I have seen further it is by standing on the shoulders of giants*^{xxii} » affirmait avec humilité Sir Isaac Newton. Propos qui expriment magnifiquement cette notion d'évolution traversée, mue par une forme de continuum invisible, de cette continuité chère à la topologie. Des épaules, des appuis, des jalons, charpente patiemment bâtie d'une œuvre collective. Perché sur les épaules d'un tiers, on voit plus loin, on est plus haut, aussi. Mais il faut s'y hisser.



Notre société numérique repose sur les épaules de géants. L'un d'eux est l'infatigable Leonhard Euler, prolifique mathématicien suisse, « *pur-sang de l'esprit* »^{xxiii} né à Bâle le 15 avril 1707, décédé à St-Petersbourg en 1783^{xxiv}, auquel la théorie des graphes et la topologie sont redevables^{xxv}.

A dire vrai, cette parenthèse historique est également motivée, un peu, par la fierté patriotique que je ressens face à d'illustres concitoyens – Leonhard Euler, Jacob Bernoulli et ses variables éponymes^{xxvi}, Ella Maillart^{xxvii}, Karl F. Jung, Roger Federer, le récent prix Nobel de chimie Jacques Dubochet^{xxviii}, et l'apatride devenu suisse : Albert Einstein^{xxix}, exemples extraits d'une liste bien fournie. Ils représentent autant de modèles dont le talent certes, mais le travail, la persévérance, l'opiniâtreté déployés au service de leurs dons, de leur curiosité, sont inversement proportionnels à l'exiguïté du territoire suisse.

Revenons à notre sujet. La théorie des graphes et la topologie possèdent une histoire étroitement liée^{xxx}, lien incarné par Euler.

La formule des polyèdres

René Descartes (1596-1650), philosophe français dont le célèbre « *cogito ergo sum* » a profondément imprégné notre société occidentale, mais également mathématicien et physicien^{xxxi}, avait relevé en 1639 un curieux schéma numérique se vérifiant « *pour les solides de toute forme, réguliers ou pas. Si le solide possède F Faces, A arêtes et S sommets alors on constate que $F - A + S = 2$* »^{xxxii}.

La formule des polyèdres^{xxxiii} nous dit que « *le nombre de faces, d'arêtes et de sommets d'un solide ne sont pas indépendants, ils entretiennent au contraire une relation très simple* »^{xxxiv}.

Prenons une forme usuelle, le cube, à l'instar de ce petit carré blanc qui adoucit nos débuts de journée ou de son cousin de glace qui fait nos délices lors d'une chaude journée d'été : 6 faces (F), 12 arêtes (A), 8 sommets (S), soit $6 - 12 + 8 = 2$. Le dodécaèdre, l'icosaèdre, le tétraèdre ou encore l'octaèdre confirment ce rapport, vous pouvez vérifier^{xxxv}. Je n'avais jamais perçu un carré de sucre ou un glaçon comme un vecteur d'harmonie dans notre monde tourmenté. Et vous ? C'est apaisant, non ?



Descartes, ne percevant dans ce constat qu'une simple particularité, n'a pas jugé utile de publier cette formule des polyèdres ; ce que fera Leonhard Euler en 1750-1751, démonstration à l'appui^{xxxvi}.

Et pourtant, ô combien ce curieux schéma s'est-il avéré être important. *Primo*, il constitue le premier *invariant topologique*^{xxxvii}, soit « une caractéristique conceptuelle qualitative d'une forme »^{xxxviii}, offrant ainsi la possibilité de distinguer entre solides de topologie différente^{xxxix}.

Secundo, il représente un jalon de l'évolution des sciences, origine^{xl} de ce qui deviendra plus tard une branche des mathématiques – la topologie^{xli} – et une *success story*^{xlii}. Et par ricochet, joue un rôle dans l'avènement et l'évolution de notre société numérique.



Topologie et graphe, la constante de la formule d'Euler « est maintenant connue comme la caractéristique d'Euler pour un graphe (ou pour un autre objet mathématique), et est liée au genre de l'objet »^{xliii}.

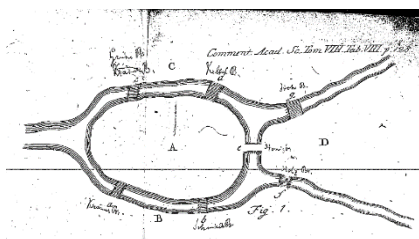
Ce n'est pas tout.

La promenade des sept ponts de Königsberg

Avant cette démonstration, Euler avait déjà posé un autre jalon, fondateur, dans l'histoire de la théorie des graphes et de la topologie en résolvant, le *Problème des sept ponts de Königsberg*^{xliv}. Il présenta pour la première fois son analyse le 26 août 1735 à l'Académie des sciences de St-Petersbourg. Elle sera publiée ultérieurement sous le titre *Solutio problematis ad geometriam situs pertinentis*^{xlv}. En français : *La solution d'un problème lié à la géométrie de situation*^{xlvi}.

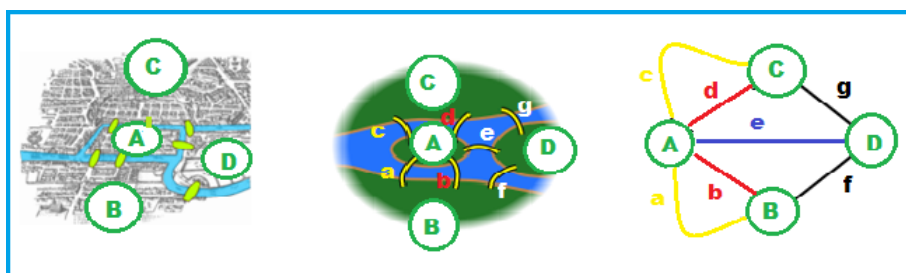
Ce casse-tête a pour cadre la ville prussienne de *Königsberg*, connue aujourd'hui sous le nom de Kaliningrad^{xlvii} : une ville, un fleuve, des ponts. Une ville ordinaire.

Le problème posé consistait à déterminer s'il était possible de prévoir une promenade à travers les rues de la ville en empruntant chacun des sept ponts une fois, et une seule fois^{xlviii}.



Un schéma tracé de la main d'Euler illustre l'approche originale du scientifique^{xlix}. Euler réduit Königsberg à quatre régions et sept ponts. Nous verrons que la résolution de ce problème réside dans cette extrême simplification de la situation géographique.

Königsberg, ville construite autour de deux îles (A, D) situées sur le fleuve Pregel partagé en deux bras. Deux îles reliées par un seul et unique pont (e). Les rives du fleuve (les régions B et C) sont rattachées à l'une ou l'autre des deux îles



par six ponts au total (a – d, f, g)^l.

Euler s'appuie sur une notion empruntée à Gottfried Leibnitz^{li}: *Geometria situs*, la géométrie de position^{lii}. Quand bien même il n'était pas clair en quoi elle consistait^{liii}, conscient de la nouveauté de la démarche^{liv}, Euler a vu dans le casse-tête qui s'offrait à lui, un cas possible d'application de la géométrie de position : « *Lorsque récemment il fut question d'un problème qui semblait, à la vérité, se rattacher à la géométrie ordinaire, mais dont cependant la solution ne dépendait, ni de la détermination de grandeurs, ni du calcul de quantités, je n'ai point balancé à le rapporter à la géométrie de position, d'autant plus que les considérations de positions entre seules dans la solution, tandis que le calculé n'y est pour rien. J'ai donc cru utile d'exposer ici, comme un exemple de géométrie de position, la méthode que j'ai trouvée pour résoudre les problèmes de ce genre* »^{lv}.

Bien sûr, précise-t-il, « *on pourrait résoudre le problème en faisant l'énumération complète de toutes les manières de passer qui peuvent avoir lieu* »^{lvi}, mais ô combien de combinaisons cela représenterait ! Travail laborieux, non réalisable quand le nombre de ponts augmente.

Il adopte une autre approche, innovante. Le choix d'un parcours à l'intérieur d'une région n'était pas pertinent^{lvii}. Des séquences de lettres en guise de notation d'un passage sur un pont : « *Toute ma méthode se fonde sur une manière particulière de représenter chaque passage de pont, dans laquelle j'emploie les lettres majuscules A, B, C, D, qui sont écrites à chaque région que sépare le fleuve. Ainsi, si quelqu'un va de la région A à la région B, en passant sur le pont a ou sur le pont b, je désigne ce passage par les lettres AB. La première marque la région d'où sort le voyageur ; la seconde, la région dans laquelle il est parvenu après avoir passé le pont. Si ensuite le voyageur s'en va dans la région D par le pont f, ce passage sera représenté par les lettres BD, et je représente ces 2 passages successifs AB et BD seulement par 3 lettres ABD, celle du milieu B représentant, tant la région où il est parvenu par un premier passage que celle d'où il est sorti pour un second passage. (...) C'est pourquoi il faut 8 lettres pour représenter les passages sur 7 ponts* »^{lviii}.

Les régions sont des sommets (nœuds), les ponts des arcs exprimés non par des lettres minuscules comme sur le schéma, mais par une suite de lettres majuscules représentant les zones géographiques^{lix}.

La démarche possède un double avantage : premièrement, résoudre le casse-tête. Euler rejette la possibilité d'une telle promenade, pour les raisons suivantes : 7 ponts. Exigence : une séquence de 8 lettres pour résoudre le cas positivement. En l'espèce : une séquence de 9 lettres. Trop de lettres. « *Dans la série des 8 lettres représentant le passage sur les 7 ponts, la*

lettre A devrait se trouver 3 fois, et les lettres B, C, D, chacune deux fois, ce qui, dans une série de 8 lettres, est complètement impossible »^{lx}.

Deuxièmement, la démarche choisie a permis à Euler de généraliser le casse-tête en posant une règle applicable à tout problème de même type^{lxi}. Il formule que l'existence de la promenade souhaitée via *les sept ponts de Königsberg* aurait requis la réalisation d'une condition relative au nombre de ponts appartenant à chaque zone géographique : a) Chaque portion de terre constitue une zone géographique (ou région) ; b) chaque région possède un degré, à savoir le nombre de ponts qui la touche ; c) le nombre de régions possédant un degré impair s'élève à 0 ou 2. Ainsi, en présence de plus de deux régions ayant un degré impair, soit lorsque plus de deux zones géographiques sont touchées par un nombre impair de ponts, le problème n'a pas de solution. Inversement, si au plus deux régions possèdent un degré impair, une solution à la promenade telle que déterminée existe forcément^{lxii}. Nécessaire, cette condition s'avère également suffisante^{lxiii}.

Cette règle revêt un caractère général^{lxiv}.

Dans le cas de *Königsberg*, la condition des degrés impairs n'est pas réalisée. Les quatre régions déterminées par Euler comme constituant la ville possèdent des degrés impairs : A, 5 degrés ; B, 3 degrés, C, 3 degrés ; D, 3 degrés. Quatre régions, c'est trop. Il en faut deux au maximum par situation !

Du croquis au graphe. Les fondations de la théorie de graphes et de la topologie étaient posées^{lxv}. Bien sûr « *Euler n'a pas tracé de graphes, il n'a pas parlé de sommets ou d'arcs ni de degré d'un sommet. Dans sa résolution du problème Euler ne trace que le schéma de la ville et ensuite il réfléchit sur ce schéma exactement comme si c'était un graphe formé de sommets et d'arcs* »^{lxvi}

Hissé sur les épaules de Leibniz, Euler a appliqué une approche géométrique inédite, abstraite, sans mesure ni détails, basée sur des connections, des liens, « *une géométrie qui ne tient pas compte des mesures (« grandeurs » ou « quantités », pour reprendre les termes d'Euler lui-même) mais seulement des positions relatives des objets considérés* »^{lxvii}. « *L'article d'Euler a été le premier qui portait spécifiquement sur cette nouvelle discipline qui allait engendrer au 19^e siècle non seulement la théorie des graphes, mais également l'analyse de position qui a mené à la topologie*^{lxviii}, *un sujet mathématique qui a connu un développement considérable au 20^e siècle* »^{lxix} pour devenir au fil des siècles un domaine important de la recherche mathématique, avec des applications dans de nombreux domaines (en chimie, biologie, neuroscience, en recherche opérationnelle, en sciences sociales et en informatique, etc.)^{lxx}.

Pour qui souhaiteraient des explications supplémentaires liées à la résolution de ce casse-tête, les références abondent, je vous en propose quelques-unes en note de bas de page, à commencer par la lecture de Leonhard Euler lui-même^{lxxi}.

En s'appuyant sur les travaux de Descartes et Leibniz, Leonhard Euler a résolu *Le problème des sept ponts de Königsberg* et énoncé la formule des polyèdres, les premiers théorèmes de la théorie des graphes et de la topologie. Durant les siècles suivants, d'autres géants se sont appuyés sur les épaules de l'illustre mathématicien suisse, ont développé la topologie et la théorie des graphes. Je clos ici cette parenthèse historique et ne m'arrêterai donc pas sur cette

période intermédiaire^{lxxii}, mais ferai directement un pas (de géant !) vers des temps qui sont les nôtres et poursuivrai notre marche vers le point de rencontre des mathématiques et de l'informatique : le moteur de recherche Google. Encore lui ! Il nous attend en seconde partie de ce Billet #4B : *Graphe et Big Data*, mais d'abord creusons la notion de graphe.

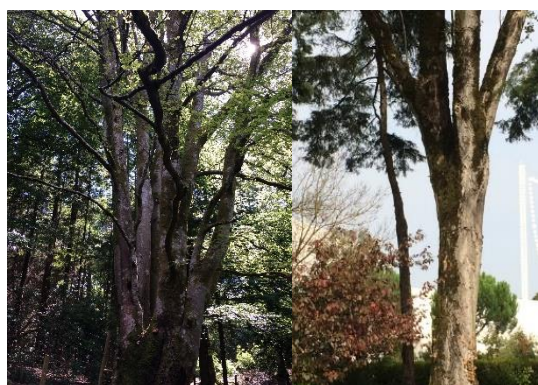
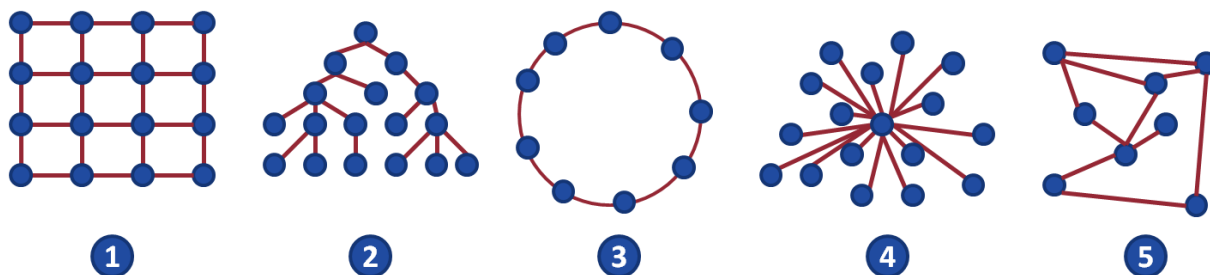
Les graphes, organiser des relations complexes aux formes variées.

Pour exploiter les données de types variés typiques du Numérique, il faut au préalable les structurer, y mettre de l'ordre^{lxxiii}. Une façon de procéder consiste à « *représenter par des graphes les données et les relations qui les lient* »^{lxxiv}. Relation, une notion intrinsèquement associée aux graphes : « *Les graphes représentent toutes sortes de relations entre des objets, partout où existe une séquence ordonnée, une dépendance temporelle ou une causalité* »^{lxxv} ; cela explique l'omniprésence des graphes, simples ou complexes : GPS, menu téléphonique, menu des applications interactives, formules chimiques, recettes de cuisines^{lxxvi}, génétique^{lxxvii}, etc.

La théorie des graphes pourrait-elle également s'appeler la théorie des relations ?

Nœud papillon, arbres, et les autres. Tant de formes insoupçonnées.

Nous avons vu au début de ce Billet #4B en mots et en substance ce que sont la théorie des graphes et la topologie. Voici maintenant à quoi ressemblent visuellement, les principales topologies typiques de graphes^{lxxviii} :



On reconnaît bien sous chiffre 2, la structure arborescente dont la forme évoque effectivement celle d'un arbre. Ces structures, sous-catégorie plus simple de graphes^{lxxix}, qui permettent d'organiser et de retrouver les données plus rapidement qu'avec d'autres techniques de stockage, sont couramment utilisées dans les sous-programmes de recherche et de tri^{lxxx}.

Au-delà des arborescences, se trouvent les graphes, un graphe se définissant également comme constituant « *une sorte d'extension du concept d'arborescence* »^{lxxxi}, utile dans différentes situations – la cartographie GPS, par exemple – ne pouvant être modélisées par une approche en arborescence du haut vers le bas^{lxxxii}.

Des perceptions différentes

Le GPS, compagnon quotidien de nos déplacements, servira d'illustration à ce paragraphe.



Nous avons déjà évoqué au Billet #4A des différences de fonctionnement fondamental entre un humain et un ordinateur, conséquence d'une approche binaire limitée à des 0 et des 1^{lxxxiii}. Voici une autre distinction.

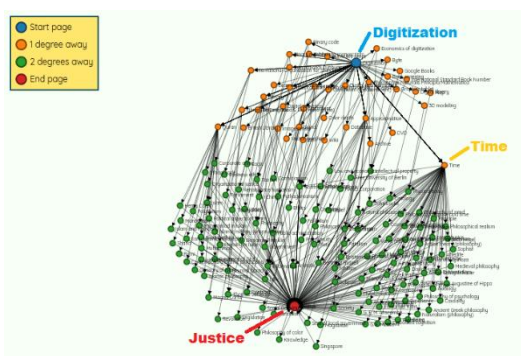


Là où nous voyons un plan, l'ordinateur perçoit un graphe. L'ordinateur n'a pas besoin d'une représentation graphique, qu'il ne comprend de toute manière pas. Cette visualisation graphique nous permet en revanche, à nous êtres humains, d'embrasser en un regard les éléments constitutifs de la carte routière ou du plan^{lxxxiv}. Par ailleurs, il s'agit d'un exemple extrêmement parlant en matière de graphes : les intersections sont les sommets (nœuds), et les rues les arêtes ou les arcs^{lxxxv}, ou à plus grande échelle : les villes sont des sommets, les autoroutes les arêtes^{lxxxvi}, par exemple.

Les graphes, organisation de l'information de liens complexes.

Les graphes constituent une structure organisationnelle contenant différents éléments informationnels :

- **Complexité.** Outils de modélisation de la complexité, les graphes composés de nœuds possédant souvent de multiples connexions, permettent de représenter des liens entre des sommets, facilitant de la sorte la détermination des relations complexes^{lxxxvii}. Il existe des graphes simples, d'autres complexes, voire extrêmement complexes. La complexité se détermine en fonction du nombre d'arêtes et de sommets^{lxxxviii}.



Voici, par exemple, un graphe complexe réalisé avec un outil amusant, *Six Degrees of Wikipedia*^{lxxxix}, qui représente les chemins les plus courts^{xc} entre les termes *Digitization* et *Justice* dans Wikipédia. Il a trouvé en 2.58 secondes (!) 169 chemins avec 3 degrés de séparation entre *Digitization* (page de départ) et *Justice* (page d'arrivée).

Une étape (nœud) qui se démarque dans le graphe entre *Digitization* et *Justice* est *Time*... *Le temps*... Le temps qui semble se raccourcir à l'ère numérique. Le rôle de ce phénomène en lien avec la justice, qui fonctionne sur autre cadence que le Numérique est une thématique que j'ai abordée dans le Billet #3D^{xc}.

- **Directionnalité.** « Contrairement à un arbre, qui est constitué de relations parent/enfant, un graphe peut comporter des liens orientés dans un sens ou dans l'autre, ou non, de la même manière que dans une agglomération la plupart des rues sont bidirectionnelles, mais certaines sont à sens unique »^{xcii}.

- On parle de *graphe non orienté* lorsque l'ordre des entrées n'est pas pertinent.



Sur un plan, cela équivaut à une route bidirectionnelle. L'arête est représentée par un trait.

- Dans un *graphe orienté*, l'ordre des entrées importe, « car le flux transite de la première entrée vers la deuxième. Dans ce cas, on parle d'arcs plutôt que d'arêtes »^{xciii}. L'arc est symbolisé par une flèche pointant dans l'unique direction. Dans le domaine routier, la séquence de feux signalétiques peut être représentée par un *graphe orienté*, l'ordre des entrées étant essentiel pour éviter le chaos : « Rouge = 1, Orange = 2 et Vert = 3. Les trois arcs nécessaires pour représenter cette séquence sont : Passez = [Rouge, Vert], Attention [Vert, Orange], et Stop = [Orange, Rouge] »^{xciv}. Un graphe orienté peut en outre être *cyclique* ou *acyclique*, selon que « les arêtes forment une boucle qui ramène au sommet initial après la visite des sommets intermédiaires »^{xcv} ou, au contraire, que « le graphe ne comporte aucune boucle »^{xcvi}. Boucle ou pas boucle, telle est la question ; une série d'étapes, qui se répète, ou pas^{xcvii}.



- Enfin, dans les cas mixtes, on parlera de *graphe partiellement orienté* ; cela correspond, dans notre exemple routier, à la présence tant de rues

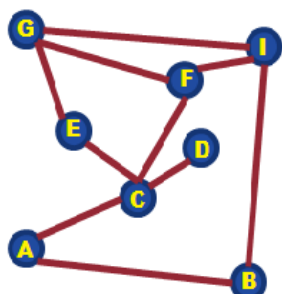


bidirectionnelles qu'à sens unique (unidirectionnelles)^{xcviii}.

- *Pondération*. Un élément de précision. « Dans un graphe, la représentation d'un lien ne reflète pas nécessairement toute la réalité. A certains liens, il est possible d'associer une pondération représentant par exemple la distance entre deux points, ou le temps nécessaire pour effectuer un parcours, etc. »^{xcix}. Dans ce contexte, on parle de graphe pondéré, soit « un graphe avec des valeurs assignées aux arêtes ou aux arcs »^c. Lors de nos déplacements, s'il s'avère essentiel de connaître la direction, la distance ou le temps de trajet constituent également des éléments appréciables, fournis par un *graphe pondéré*.
- *Centralité*. La détermination et la visualisation de la centralité permet de déterminer les « sommets les plus importants du graphe (...) savoir quelles parties exigent la plus grande attention »^{ci}. Nous retrouverons cet élément informationnel ci-après dans le chapitre *PageRank*^{cii}.
- *Fonctionnalité*. Soit pour un graphe la « capacité à modéliser un problème donné »^{ciii}.

Plus formellement.

Formulé de manière plus scientifique, « un graphe est constitué de paires ordonnées sous la forme $G = (V, E)$, où G est le graphe, V est une liste de vertex, c'est-à-dire de sommets, et E une liste d'arêtes reliant les sommets. Une arête est en réalité une paire numérique représentant une liaison entre deux sommets »^{ciiv} ; à n'en pas douter, Euler sourit en lisant ces lignes.



Reprenons le graphe 5 ci-dessus, complété par des lettres représentant des villes (sommets), reliées par des autoroutes (arêtes). Cette carte routière constitue *un graphe non orienté* puisque chaque autoroute est à double sens. Si nous nous penchons, par exemple, sur le trajet entre A et B, cela nous donne : Autoroute = [A, B], avec le graphe noté $G = [(A, B)]$. On comprend ainsi qu'il est possible de se rendre de la ville A à la ville B en empruntant l'autoroute^{cv}.

Les incontournables : les réseaux.

La notion de relations, sous une forme ou une autre, est indissociable de la notion de graphe^{cv}. Faisons un pas de plus. Les graphes nous rapprochent d'une notion devenue omniprésente à l'ère numérique : les réseaux. La notion de réseau n'est pas récente, présente dans de nombreux domaines : réseaux électriques, réseaux neuronaux, réseaux de transports, réseau routier, réseau de communication, etc. ce qui est nouveau ce sont « *these new human-machine networks* »^{cvii} à l'échelle mondiale. *Les réseaux sociaux.*

De manière générale, Wikipédia définit de la sorte les réseaux : « *En premier ressort, un réseau désigne au sens concret « un ensemble de lignes entrelacées » et, au figuré « un ensemble de relations ».* Par extension, il désigne un ensemble interconnecté, fait de composants et de leurs inter-relations, autorisant la circulation en mode continu ou discontinu de flux (eau, air, huile...) ou d'éléments finis (marchandises, informations, personnes...). Le réseau peut être « matériel » (comme le réseau électrique, le réseau routier, le réseau sanguin ou le réseau lymphatique), « immatériel » (comme le réseau social), « abstrait, symbolique ou normalisé » (comme le réseau de tâches de la méthode PERT) »^{cviii}.

John Paul Mueller et Luca Massaron, auteurs de l'ouvrage *Les algorithmes pour les nuls*, proposent une définition autour du graphe, ainsi : « *Un réseau est un type de graphe associant des noms à des sommets (ou nœuds) et/ou des arêtes (ou arcs, ou lignes).* L'utilisation de noms rend la représentation graphique moins abstraite et facilite la compréhension. Le lecteur a ainsi une vision plus concrète des données représentées, même si le graphe reste une représentation abstraite de la réalité sous une forme pouvant être comprise de façon différente par l'utilisateur et par la machine »^{ciix}.

Ainsi, réseaux et théorie des graphes font bon ménage : « *Les algorithmes élaborés pour résoudre des problèmes concernant les objets de cette théorie ont de nombreuses applications dans tous les domaines liés à la notion de réseau* »^{cx}.

Le Web, réseau des réseaux.

« *Toile (d'araignée) mondiale* »^{cx}, graphe en forme de nœud papillon. Mère Nature nous prête ses habitants, précieuses métaphores offrant une représentation différente du Web : « *Le World Wide Web, en tant qu'ensemble de ressources hypertextes, est modélisable en graphe orienté possédant des cycles* »^{cxii} avec les ressources pour sommets et les hyperliens pour arcs.

Comme le graphe est orienté, certaines ressources peuvent constituer des puits, ou moins formellement des culs-de-sac : il n'existe aucun chemin vers le reste du web. À l'inverse, certaines ressources peuvent constituer des sources : il n'existe aucun chemin depuis le reste du web »^{cxiii}.

Les auteurs de la recherche *Graph structure in the web*, évoquée ci-dessus en introduction de du chapitre *Le Web, un graphe de forme familière*, ont considéré un graphe orienté dont les nœuds correspondaient aux pages statiques du Web, et les arcs aux liens entre les pages^{cxiv}, graphe dont ils ont ensuite analysé, décortiqué les propriétés^{cxv} pour conclure que « *Our study indicates that the macroscopic structure of the web is considerably more intricate than suggested by earlier experiments on a smaller scale* »^{cxvi}. Une étude à grande échelle qui a également mis à jour la complexité de la structure du Web.

Le Web c'est donc une forme en nœud papillon, les nœuds peuvent être des pages, mais les nœuds des réseaux sociaux peuvent être aussi des personnes « *thinking human beings. They can make decisions, potentially changing their networks even while embedded in them and being affected by them. A network of humans has a special kind of life of its own* »^{cxvii}. Ici, comme souvent, le meilleur comme le pire s'offrent à nous. Notre liberté et son corollaire, notre responsabilité, sont convoquées. Qu'allons-nous en faire ?

D'abord objets d'une discipline uniquement mathématique, les graphes cohabitent désormais fructueusement avec l'informatique. La rencontre des mathématiques (via la théorie des graphes et de la topologie) et de l'informatique (à la lumière de l'évolution du moteur de recherche de Google, et de Hadoop), a fait du graphe l'une des pierres angulaires du Big Data^{cxviii}.

GRAPHE ET BIG DATA

Nous poursuivons notre marche vers le moteur de recherche Google : point de rencontre des mathématiques et de l'informatique et sommet de ce Billet. Toutefois, avant cela, une étape préalable nous attend, étape au paysage familier : *Hadoop*, que nous contemplant cette fois chaussés des lunettes de la théorie des graphes et de la topologie.

Hadoop, la topologie et les graphes.

Le terme topologie a débordé ses frontières initiales pour s'étendre à l'informatique. La topologie, en mathématiques, c'est avant tout une question de formes, d'espaces. En informatique, ce n'est pas différent. Par *topologies*, on entend les différentes formes que peuvent prendre des réseaux^{CXXIX}, d'où l'expression suivante : *topologie de réseau*, qui se réfère « à l'architecture (physique ou logique) de celui-ci, définissant les liaisons entre les équipements du réseau et une hiérarchie éventuelle entre eux. Elle peut définir la façon dont les équipements sont interconnectés et la représentation spatiale du réseau (topologie physique). Elle peut aussi définir la façon dont les données transitent dans les lignes de communication (topologies logiques) »^{CXX}. La topologie physique décrit donc la structure physique du réseau, sa forme, son apparence ; elle nous informe sur la manière dont les ordinateurs sont reliés entre eux. La topologie logique, constitue la structure logique d'une topologie physique et elle définit comment s'y déroule la communication^{CXXI}.

Réseau d'ordinateurs, architecture, communication. Voici des concepts familiers. Effectivement, dans la mini-série de Billets #3 nous avons abordé sans le savoir des éléments de topologie informatique. : clusters, architectures distribuées^{CXXII}. Ou encore stockage des données distribué (*Hadoop Distributed System File*, HDSF) et traitement des données massivement parallèle (MapReduce) sur les nœuds d'un *cluster en architecture shared-nothing* constitué de nœuds indépendants qui ne communiquent pas entre eux. Un orchestrateur des tâches se trouve en charge de la communication dans ce type cluster^{CXXIII}.

Sans le savoir, nous avons également touché à la théorie des graphes. En informatique, un graphe se définit comme le plan d'exécution d'une requête ou d'un programme par un moteur de calcul^{CXXIV}. En d'autres termes, il s'agit du descriptif de l'enchaînement des opérations nécessaire pour qu'un programme réponde au besoin pour lequel il a été conçu. Formellement, un graphe consiste en un ensemble de sommets (ou vertices ; des rectangles pour les fichiers de données, des cercles pour les traitements de données) et de flèches (les arcs). Le MapReduce est un exemple de graphe composé de trois sommets - le Map, le Shuffle et le Reduce – et qui s'exécute en suivant un schéma acyclique orienté^{CXXV}. Il s'exécute systématiquement selon un schéma identique : Map -> Shuffle -> Reduce^{CXXVI}.

Promenons-nous dans le Web.



Quel bazar, n'importe qui s'y perdrait ! A l'instar de cette bibliothèque, le Web constitue un vaste monde^{CXXVII}, composé de lieux accessibles, mais aussi de contrées reculées, difficilement atteignables. Comment s'orienter dans cet espace aux millions de propositions, étale aux mille saveurs ? Comment y trouver la perle, la bonne information ?

Améliorer le classement pour trouver la bonne information

Les moteurs de recherche, sésames qui nous ouvrent l'accès au Web, nous permettent de dénicher l'information désirée en une simple requête, de faire notre marché dans ce vaste espace encombré ; voilà pour le côté scène. En coulisses, se déroule un travail d'indexation^{CXXVIII}, condition préalable mais non suffisante à une recherche fructueuse : « *La correspondance entre les mots, entre la requête et le texte de la page (ou la correspondance sémantique (...)), est une condition préalable, mais non suffisante, car les hyperliens sont nécessaires pour déterminer si la page présente un contenu de qualité et constitue une référence sur un sujet concerné* »^{CXXIX}. Les hyperliens jouent un rôle primordial dans la recherche de la bonne information, ils « *sont nécessaires pour déterminer si la page présente un contenu de qualité et constitue une référence sur le sujet concerné* »^{CXXX}. Plus il y a d'hyperliens conduisant sur une page, plus l'algorithme considère la page comme importante (« relevant »).

Tout évolue. « *Avant Google, le web était une vaste loterie* »^{CXXXI}. Avec les premiers moteurs de recherches, le classement était simpliste, souvent fondé sur la fréquence des mots-clés sur les pages, sur leur apparition dans le titre ou l'en-tête des pages. Les invariants n'étant pas l'apanage de la topologie, présents également dans la nature humaine, se manifestant sous des traits chatoyants, moins reluisant également à l'instar du corbeau ou du singe des fables de la Fontaine (un autre géant), d'aucuns ont su utilement exploiter insidieusement les faiblesses d'enfance de ces techniques, pour en tirer des avantages indus en truquant les résultats des recherches, un bon classement attribué à une page au contenu mauvais, voire trompeur : les spammeurs du Web, les « chapeaux noirs »^{CXXXII}.

L'émergence de ces nouveaux acteurs indéliçables, la possibilité de manipuler les moteurs de recherche, la structure toujours croissante du Web, la difficulté pour un utilisateur d'obtenir un résultat satisfaisant à une requête, imposaient de doter les moteurs de recherche des meilleurs algorithmes de classement.

De quelle manière améliorer le classement ?

Eurêka ! PageRank et Google.

Commençons par poser quelques jalons. Nous effectuons un voyage en Terra Data sur les épaules des géants. Notre monde numérique a aussi ses géants, les géants du Web, notamment, dont Google fondée en septembre 1998 par Larry Page et Sergueï Brin, concepteurs du fameux moteur de recherche. Dans le domaine de l'amélioration des

recherches sur le Web, comme souvent dans d'autres domaines, différents projets se sont développés en parallèle. Au final, l'algorithme conçu par Serguei Brin et Larry Page s'est imposé^{cxixiii}. Il s'appelle : *Page Rank*. Nommé d'après le patronyme de l'un de ses concepteurs – Larry Page - il fit sa première apparition publique le 29 janvier 1998 dans un article intitulé *The PageRank citation rankings : bringing order to the web*^{cxixiv}, article au titre annonciateur d'une mission herculéenne : apporter de l'ordre dans le Web.

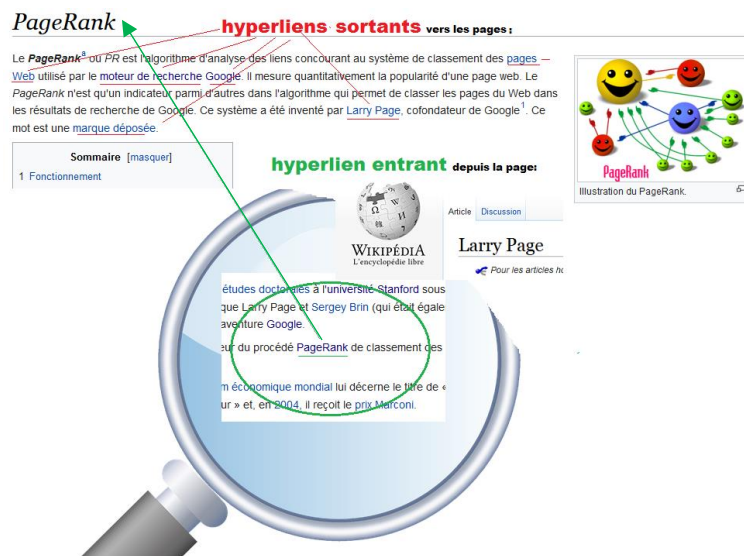
Le premier brevet de PageRank, technologie socle du moteur de recherche Google, a été enregistré en janvier 1998 alors que ses deux inventeurs étaient encore doctorants à l'Université de Stanford ; cette dernière concèdera une licence à Google, fondée quelque mois plus tard^{cxixv}.

Suivra, en avril 1998, l'annonce de la naissance d'un futur géant du Web : le moteur de recherches Google: *The Anatomy of a Large-Scale Hypertextual Web Search Engine*^{cxixvi}, publication dont voici les premiers mots : « *In this paper, we present Google, a prototype of a large-scale search engine (...)* »^{cxixvii}. Les auteurs expliquent ensuite que leur moteur de recherche est conçu pour explorer et indexer le Web efficacement et fournir des résultats de recherche beaucoup plus satisfaisants que ceux des systèmes existants^{cxixviii}. Leur idée consiste à exploiter la structure présente dans l'hypertexte^{cxixix}, à utiliser les informations supplémentaires qui s'y trouvent afin d'améliorer les résultats de recherche^{cxli}. Les informations supplémentaires ? Les liens hypertexte ou hyperliens^{cxli} car « *Le lien hypertexte équivaut à une citation et que, à sa manière, il peut être considéré comme un vote* »^{cxlii}. Il leur importait également de se pencher sur la gestion efficace des collections d'hypertextes non contrôlées où tout quidam peut publier à sa guise ce qu'il souhaite^{cxliii}.

Les hypertextes, l'ADN même du Web, « système hypertexte public fonctionnant sur Internet. Le Web permet de consulter, avec un navigateur, des pages accessibles sur des sites. L'image de la toile d'araignée vient des hyperliens qui lient les pages web entre elles »^{cxliiv}. Le Web, système de publication d'informations sous forme de documents hypertextes reliés entre eux^{cxliv} par des hyperliens, permettant aux utilisateurs d'aller directement et selon leur bon vouloir à la page qui les intéressent, d'une manière non linéaire^{cxlvi}. Les hyperliens, liens constituant des éléments de la modélisation de la Toile « *en graphe orienté possédant des cycles avec les ressources pour sommets et les hyperliens pour arcs* »^{cxlvii}. Arcs unidirectionnels vers les ressources accessibles d'Internet^{cxlviii}. D'hyperliens en hyperliens, *la Toile* s'est tissée... *La Toile... Les graphes...* Une ressemblance visuelle, mais également le point de convergence de plusieurs disciplines sous une bannière : *PageRank*

PageRank.

PageRank^{cxlix}, c'est la rencontre des mathématiques et de l'informatique, par la mise en œuvre d'un « modèle mathématique, le graphe, jusqu'alors peu utilisé en informatique »^{cl}. La mise à



profit de la structure en liens du Web pour ordonner son hétérogénéité. Il permet des références croisées, fait des recommandations, à l'instar d'une bibliothécaire à laquelle on demanderait conseil^{cli}. PageRank propose une solution aux problèmes de classement de l'information qui utilise avantageusement ces liens hypertextes^{clii}, en s'appuyant sur l'existant, le déjà connu, épaules solides pour voir plus

loin, socle pour s'élever, grimper, dévoiler de nouveaux possibles, émergeant des graines du précédent.

PageRank, c'est la mesure de la popularité d'une page Web opérée par l'attribution de « scores aux sommets d'un graphe de telle sorte que plus le score attribué à un sommet est élevé, plus le sommet est considéré comme important dans le graphe »^{cliii}. Déterminer cette importance « signifie évaluer la pertinence de chaque page des résultats de la requête, afin de mieux servir les utilisateurs qui recherchent un contenu intéressant. Une page est une bonne

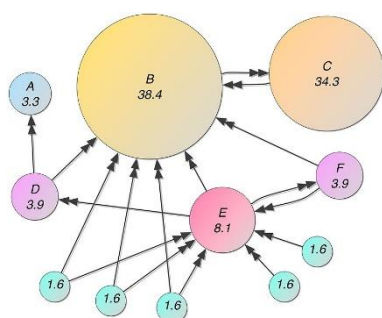


réponse à une requête si elle correspond aux critères de cette requête et si elle occupe une place importante dans le système d'hyperliens qui relie les pages les unes aux autres. Le principe est que la « toile » est ce qu'en font les utilisateurs, et que si une page est importante sur le réseau, c'est pour une bonne raison (la qualité et l'autorité du contenu de la page sont évaluées en fonction de son importance dans le réseau des hyperliens) »^{cliv}, soit du nombre d'hyperliens qui pointent dessus.

PageRang, c'est l'objectivation d'un comportement. Il peut être considéré comme un modèle du comportement d'un utilisateur^{clv}, ainsi que l'expliquent ses pères, que je traduis ici : « L'importance d'une page Web est une question intrinsèquement subjective, qui dépend des intérêts, des connaissances et des attitudes des lecteurs. Mais il y a encore beaucoup de choses que l'on peut dire objectivement sur l'importance relative des pages Web. Cet article décrit le PageRank, une méthode d'évaluation objective et mécanique des pages Web, mesurant efficacement l'intérêt humain et l'attention qui leur est accordée. Nous comparons le PageRank à un internaute aléatoire idéalisé »^{clvi}.

Mesure de centralité^{clvii} sur la Toile, le principe de base de PageRank consiste à attribuer « un score à chaque sommet d'un graphe (plus ce score est élevé, plus le sommet est important). Le

score attribué à une page Web représente la probabilité qu'un internaute aléatoire visite cette page »^{clviii}. Le score d'une page, son PageRank est une probabilité. Pour le détail du calcul de ces probabilités, je renvoie le lecteur intéressé aux publications-sources^{clix}.



Qui mieux que ses pères peut proposer une description intuitive de PageRank et de son système de liens retour (*backlinks*) pointant vers un site ou une page Web^{clx} : « *A page has high rank if the sum of the ranks of its backlinks is high. This covers both the case when a page has many backlinks and when a page has a few highly ranked backlinks* »^{clxi}.

Voici ci-contre un exemple illustré très parlant proposé sur la page *Page Rank* de *Wikipedia* en langue anglaise^{clxii}.

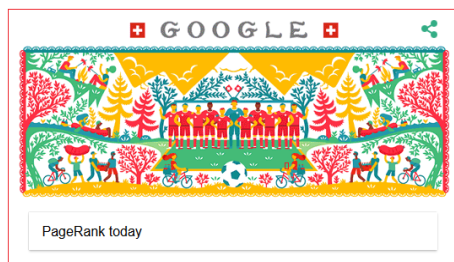
Les hyperliens, maillons d'une chaîne de confiance. Dans le monde numérique, la confiance est essentielle, tout comme elle l'est dans le monde matériel. En revanche, elle ne se construit pas de la même manière. Là où elle repose sur la proximité, le contact, la familiarité, notamment, dans le monde numérique, elle se bâtit différemment, la proximité, n'étant pas un critère-clé. « *Un hyperlien est une façon de cautionner ou de recommander une page. Les liens entrant vous montrent que les créateur de pages vous font confiance, et la confiance est partagée lorsque vous ajoutez vous-même à vos pages des liens sortants qui pointent vers les pages des autres* »^{clxiii}. « *Chaque page transfère un peu de sa popularité aux pages qu'elle référence. Une page est donc d'autant plus populaire qu'elle est référencée par de nombreuses pages populaires* »^{clxiv}.



PageRank, c'est un cœur au poulx aujourd'hui plus discret. Le fonctionnement du moteur de recherche de Google a évolué. Depuis sa version originale, l'algorithme PageRank a subi de nombreuses mises à jour : « *Google n'a pas seulement ajouté des facteurs de positionnement (...), la compagnie a réalisé d'importants changements pour améliorer l'évaluation du contenu des pages (afin d'éviter que l'algorithme soit abusé par la présence de certains mots-clés) et a adopté les algorithmes d'intelligence artificielle qui notent la pertinence d'une page dans les résultats de recherche de façon autonome* »^{clxv}. PageRank, ne constitue donc pas la seule composante de Google ; 200 autres facteurs contribuent aux résultats fournis par le moteur de recherche^{clxvi}. Une évolution significative, des algorithmes d'intelligence artificielle ont été intégrés^{clxvii} : *RankBrain*^{clxviii}, par exemple, incorporé en 2015 et fortement lié à *Hummingbird* (2013)^{clxix}. Avec cette technologie, on touche à la sémantique, à la compréhension par la machine du sens des requêtes exprimées en langage naturel par les utilisateurs, on va désormais au-delà de chaînes de mots-clés : « *Cette technologie ne s'applique pas à l'analyse des pages web, mais avant tout à établir une meilleure compréhension de ce que désire l'internaute, à extraire le concept demandé de la demande émise par l'utilisateur. Et à lui fournir par la suite la réponse la plus pertinente possible...* »^{clxx}. « *RankBrain is unique because, when it sees an unfamiliar word or phrase it uses AI to instantly search terms that have a similar meaning that a user has typed. Next, it filters the result accordingly* »^{clxxi}. La nuance languagière s'invite grâce à la compréhension des synonymes, des concepts, du mode conversationnel, de l'argot, de la sémantique des requêtes et du contenu des pages, des

ambiguïtés, des termes peu clairs, familiers. Indépendamment de l'aide de PageRank, RankBrain est capable d'analyser la façon dont une page est écrite et d'estimer si son contenu justifie qu'elle soit incluse dans les résultats de la requête^{clxxii}.

L'adoption de l'intelligence artificielle a fait émerger la question controversée de savoir si PageRank détermine toujours la position d'une page dans les résultats d'une requête. De l'avis de John Paul Mueller et Luca Massaron, auteurs de l'ouvrage *Les algorithmes pour les nuls*: « On peut supposer que PageRank continue de jouer un rôle important dans le moteur de recherche de Google en tant que facteur de positionnement, même s'il ne peut plus à lui seul déterminer la présence d'une page dans les meilleurs résultats d'une requête »^{clxxiii}.



Si vous êtes curieux d'en savoir plus sur l'état des lieux de PageRank à ce jour, je vous laisse de soin de taper « PageRank today » dans votre moteur de recherche préféré.

Défini de manière générale, « PageRank est une fonction qui attribue un score à chaque sommet d'un graphe (plus ce score est élevé, plus le sommet est élevé) »^{clxxiv}. En effet, « les mathématiques du PageRank sont entièrement générales et s'appliquent à n'importe quel graphique ou réseau dans n'importe quel domaine »^{clxxv}. « Il existe un certain nombre de versions différentes de PageRank, la recette changeant un peu à chaque fois pour s'adapter au graphe à traiter »^{clxxvi}. Dès lors, l'utilisation de PageRank ne se limite pas au moteur de recherche Google. Son application peut être envisagée « chaque fois que le problème à résoudre peut être présenté sous forme de graphe »^{clxxvii}. PageRank trouve application en lien avec Internet, ou hors Internet. Un billet intitulé *More than just a Web Search algorithm: Google's PageRank in non-Internet contexts*^{clxxviii} publié sur le site de l'Université Cornell répertorie quelques exemples de son usage dans le domaine du sport, de la littérature, des neurosciences, de la gestion des déchets toxiques, du débogage informatique, de la prévision du trafic routier et piéton dans les espaces urbains. Perdu, on lui doit de retrouver sereinement notre route grâce au GPS de notre voiture, doté d'un classique de l'algorithmie pour nous indiquer le chemin le plus court jusqu'à notre destination^{clxxix}. Certains algorithmes portent les noms de leurs auteurs, comme une rue ou une place : algorithmes de Dijkstra, de Bellman-Ford ou encore de Floyd-Warshall, s'agissant de résoudre le problème du chemin le plus court^{clxxx}. Le chemin le plus court entre deux points n'est pas forcément la ligne droite ; cela vaut pour les déplacements aidés d'un GPS mais également pour un réseau de lignes électriques^{clxxxi}. Dans le domaine de la mobilité, la théorie des graphes permet également d'analyser le trafic afin de l'améliorer, ainsi que le réseau des rues proprement dites. Lesquelles pourraient être utilement élargies, lesquelles subissent un important trafic et dès lors devront subir un entretien plus fréquent, etc^{clxxxii}. Les systèmes de menu téléphonique constituent également une forme de graphes^{clxxxiii}. On peut ajouter la biologie computationnelle, la détection des fraudes, la recommandation de produits^{clxxxiv}. Et plus encore selon Wikipedia : « Ainsi, PageRank est désormais régulièrement utilisé dans la bibliométrie, l'analyse des réseaux sociaux et d'information, ainsi que pour la prédiction et la

recommandation de liens. (...). Un PageRank personnalisé est utilisé par Twitter pour présenter aux utilisateurs d'autres comptes qu'ils peuvent souhaiter suivre »^{clxxxv}.



Twitter, et ces logos, si familiers. Les interactions sur les réseaux sociaux sont également analysées à l'aide de graphes, modélisation des données réelles. De nos données^{clxxxvi}. Sur tout sujet. Vous souvenez-vous du graphe complexe élaboré par *Six Degrees of Wikipedia*^{clxxxvii} ?

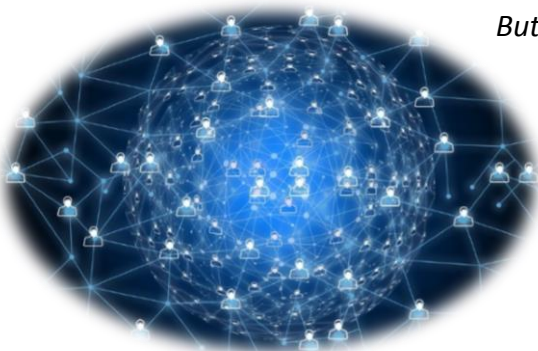
CONCLUSION

Rechercher une information sur Google fait à ce point partie intégrante de nos habitudes que cela a donné un verbe : *googliser*^{clxxxviii}.



On peut donc dire que *PageRank is almost everywhere*.

PageRank, c'est « une des pierres angulaires du Big Data, sans en être toutefois à lui seul la notion principale »^{clxxxix}.



But, graphs are everywhere ; cela même si nous ne décelons pas toujours ces objets relationnels, tapis en filigrane de notre monde numérique.



Le graphe, *Page Rank* : pierres angulaires du Big Data. Nous en avons déjà rencontré une autre : le parallélisme^{cxc}.

La théorie des graphes permet d'organiser, de structurer des données. De modéliser un problème particulier. D'extraire du sens des données, des structures utiles, des informations latentes, cachées, non évidentes à déceler. Des liens, des regroupements, des communautés. Extraire de l'information, de la connaissance. De la richesse, aussi. Les données – pétrole du XXI^e siècle.



Un géant, Leonhard Euler - génie scientifique à la curiosité multiple et généreux pédagogue - a sa vie durant semé des graines de connaissance. Elles ont germé, développé des racines et crû au fil des siècles suivants. Elles portent aujourd'hui des fruits variés, qui alimentent nos quotidiens numériques.



Quel chemin parcouru depuis *les sept ponts de Königsberg* !

Ainsi s'achève ce Billet #4B consacré à l'apport des mathématiques dans le développement des moteurs de recherche et du Big Data.

A bientôt pour le Billet #4C !

Anne-Sylvie Weinmann

Références :

- ⁱ Pour quelques statistiques liées à Internet et au Web voir : *Internet live stats* (<http://www.internetlivestats.com/>). « *The Indexed Web contains at least 4,39 billion pages* » (05/09/2018 ; <http://www.worldwidewebsize.com/>). Remarquez-vous la chute drastique du nombre de pages Web indexées au printemps ?! Effet RGPD ? Puis le nombre est remonté. La même référence, dans mon Billet #3C mentionnait 4,49 milliards de pages indexées en date du 20/11/2017.
- ⁱⁱ Premier travail sur le sujet : BRODER Andrei, KUMAR Ravi, MAGHOUL Farzin, RAGHAVAN Prabhakar, RAJAGOPALAN Sridhar, STATA Raymie, TOMKINS Andrew, WIENER Janet, *Graph Structure in the web*, Computer Networks, Volume 33, Issues 1–6, June 2000, 309-320. Référence du schéma : Figure 9, p. 11 de l'article en version pdf ([https://doi.org/10.1016/S1389-1286\(00\)00083-9](https://doi.org/10.1016/S1389-1286(00)00083-9), disponible sur : <http://www.immorlica.com/socNet/broder.pdf>). Publication ultérieure sur le même sujet qui a également mis à jour une structure du Web en forme de nœud papillon : MEUSEL Robert, VIGNA Sebastiano, LEHMBERG Oliver, BIZER Christian, *Graph Structure in the Web — Revisited or A Trick of the Heavy Tail* published in : WWW '14 Companion Proceedings of the 23rd International Conference on World Wide Web,, Seoul, Korea, April 07-11, 2014, Pages 427-432 (<https://doi.org/10.1145/2567948.256928>, disponible sur : <https://planet-data.org/sites/default/files/publications/GraphStructureRevisited.pdf>).
- ⁱⁱⁱ MUELLER John Paul, MASSARON Luca, *Les algorithmes pour les nuls*, Paris, Editions First, 2017, pp. 133ss (<https://www.eyrolles.com/Informatique/Livre/les-algorithmes-pour-les-nuls-9782412025901>).
- ^{iv} Les raisons qui les ont motivés à développer une compréhension de ce graphe étaient les suivantes :
- « 1. *Designing crawl strategies on the web*
 2. *Understanding of the sociology of content creation on the web.*
 3. *Analyzing the behavior of web algorithms that make use of link information. To take just one example, what can be said of the distribution and evolution of values on graphs like the web?*
 4. *Predicting the evolution of web structures such as bipartite cores and webrings, and developing better algorithms for discovering and organizing them.*
 5. *Predicting the emergence of important new phenomena in the web graph* » (BRODER Andrei and al., *op. cit.*, Introduction, p. 1).
- ^v BRODER Andrei and al., *op. cit.*, Abstract, p. 1.
- ^{vi} *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes. Cf. infra p. 4 *La formule des polyèdres*.
- ^{vii} *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes; STEWART Ian, *op. cit.*, p. 177.
- ^{viii} STEWART Ian, *op. cit.*, p. 124. Cf. infra page 5 et note de bas de page xlii.
- ^{ix} La topologie aide à comprendre le fonctionnement de l'ADN, molécule de la vie, « *double hélice, comme deux escaliers en colimaçon enroulés l'un autour de l'autre* » STEWART Ian, *op. cit.*, p. 140 et sur le même thème pp. 121, 124, 140-141.
- ^x La topologie trouve également application dans l'étude du fonctionnement du cerveau (Interview de Kathryn HESS par MINET Pascaline, *La boss des maths*, Le Temps du 10/04/2018 (<https://www.letemps.ch/sciences/kathryn-hess-boss-maths>)).
- ^{xi} Le livret de l'exposition *Kung Fu Motion* qui s'est tenue au ArtLab EPFL d'avril à août 2018 (<https://artlab.epfl.ch/kungfu-motion>), mentionne la topologie, au chapitre Analyses Interactives (->11) : « *Les techniques d'augmentation annotent numériquement a capture de mouvement avec des visualisations qui améliorent leur potentiel communicatif et représentent de nouvelles analyses esthétiques et scientifiques des mouvements incorporés dans l'espace et le temps. La capture de mouvement offre la base sur laquelle construire un modèle topologique continu, permettant un engagement avec la qualité affective du mouvement et un contrepoint contemporain aux méthodes de notations traditionnelles* ».
- ^{xii} STEWART Ian, *op. cit.*, p. 121. Wikipedia donne la définition suivante de la topologie « *In mathematics, topology (from the Greek τόπος, place, and λόγος, study) is concerned with the properties of space that are preserved under continuous deformations, such as stretching, crumpling and bending, but not tearing or gluing. This can be studied by considering a collection of subsets, called open sets, that satisfy certain properties, turning the given set into what is known as a topological space. Important topological properties include connectedness and compactness. (...) Topology can be formally defined as "the study of qualitative properties of certain objects (called topological spaces) that are invariant under a certain kind of transformation (called a continuous map),*

especially those properties that are invariant under a certain kind of invertible transformation (called homeomorphism) » *Topology*, <https://en.wikipedia.org/wiki/Topology>

xiii *Topologie*, <https://fr.wikipedia.org/wiki/Topologie>

xiv « La topographie (du grec *topos* = lieu et *graphein* = dessiner) est la science qui permet la mesure puis la représentation sur un plan ou une carte des formes et détails visibles sur le terrain, qu'ils soient naturels (notamment le relief et l'hydrographie) ou artificiels (comme les bâtiments, les routes, etc.). Son objectif est de déterminer la position et l'altitude de n'importe quel point situé dans une zone donnée, qu'elle soit de la taille d'un continent, d'un pays, d'un champ ou d'un corps de rue » (*Topographie*, <https://fr.wikipedia.org/wiki/Topographie>).

xv Interview de Kathryn HESS par MINET Pascaline, *La boss des maths*, Le Temps du 10/04/2018 (<https://www.letemps.ch/sciences/kathryn-hess-boss-maths>).

xvi STEWART Ian, *op. cit.*, p. 123; Königsberg bridge problem, <https://www.britannica.com/science/Konigsberg-bridge-problem>

xvii BOURSIN Jean-Louis, *Les maths pour les nuls*, Paris, Editions First, 2005, p. 380.

xviii STEWART Ian, *op. cit.*, p. 124.

xix STEWART Ian, *op. cit.*, p. 123-124.

xx STEWART Ian, *op. cit.*, p. 124.

xxi *Topology*, <https://en.wikipedia.org/wiki/Topology>

xxii *Isaac Newton*, https://en.wikiquote.org/wiki/Isaac_Newton

xxiii BARILIER Etienne, *Leonhard Euler. La clarté de l'esprit*, Lausanne, Presse polytechniques et universitaires romandes, Savoir Suisse, p. 33.

xxiv BARILIER Etienne, *op. cit.*, 21 et 154 ; *Leonhard Euler*, https://fr.wikipedia.org/wiki/Leonhard_Euler (Photo).

xxv STEWART Ian, *op. cit.*, p. 124; *Problème des 7 ponts de Königsberg*, https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg

xxvi Ces variables peuvent prendre deux valeurs : 1 ou 0, succès ou échec ; *Loi de Bernoulli*, https://fr.wikipedia.org/wiki/Loi_de_Bernoulli

xxvii LAPIERRE Alexandra, MOUCHARD Christel, *Elles ont conquis le monde : les grandes aventurières 1850-1950*, Paris, Arthaud Poche – Flammarion, 2007, 2015, pp. 235 – 243.

xxviii DUBOCHET Jacques, *Parcours*, Léchelles, Rosso Editions, 2018.

xxix KLEIN Etienne, *Le pays qu'habitait Albert Einstein*, Arles, Actes Sud, 2016.

xxx CARLSON Stefan C., *Graph Theory*, Encyclopaedia Britannica (<https://www.britannica.com/topic/graph-theory>).

xxxi *René Descartes*, https://fr.wikipedia.org/wiki/Ren%C3%A9_Descartes

xxxii STEWART Ian, *op. cit.*, p. 123 et pp. 124ss pour une démonstration.

xxxiii A plusieurs faces, « un polyèdre est une forme géométrique à trois dimensions (un solide géométrique) ayant des faces planes polygonales qui se rencontrent selon des segments de droite qu'on appelle arêtes » *Polyèdre*, <https://fr.wikipedia.org/wiki/Poly%C3%A8dre>; un chapitre entier « *Beaucoup de bruit pour des nœuds .La formuler d'Euler pour les polyèdres* », lui est consacré dans STEWART Ian, *op. cit.*, pp. 121ss.

xxxiv STEWART Ian, *op. cit.*, p. 121; BARILIER Etienne, *op. cit.*, pp. 37, 67-68.

xxxv STEWART Ian, *op. cit.*, p. 123. Il existe toutefois des contre-exemples (les polyèdres-hérissos, BARILIER Etienne, *op. cit.*, p. 68). La formule de Descartes-Euler ne s'applique pas non plus à un solide ornant notre quotidien : le cadre à photos (STEWART Ian, *op.*, pp. 126ss).

xxxvi STEWART Ian, *op. cit.*, p. 124.

xxxvii STEWART Ian, *op. cit.*, p. 121.

xxxviii STEWART Ian, *op. cit.*, p. 130.

xxxix STEWART Ian, *op. cit.*, p. 121.

xl STEWART Ian, *op. cit.*, p. 121 ; « *L'étude et la généralisation de cette formule, notamment par Cauchy et L'Huilier, est à l'origine de la topologie* » in : *Leonhard Euler*, https://fr.wikipedia.org/wiki/Leonhard_Euler

xli Johann Listing a utilisé le premier le terme topologie en 1847 dans un ouvrage intitulé *Vorstudien zur Topologie*, *Etudes de topologie*, soulignant de la sorte l'autonomie croissante de cette discipline (Johann_Benedict_Listing, https://fr.wikipedia.org/wiki/Johann_Benedict_Listing; STEWART Ian, *op. cit.*, p. 129 ; *Topology*, <https://en.wikipedia.org/wiki/Topology>).

xlii « *Beaucoup de temps s'écoulerait avant que les mathématiciens voient dans cette petite équation le premier pas vers la grande success story des mathématiques du XX^e siècle, l'irrésistible ascension de la topologie. Au XIX^e siècle les trois piliers des mathématiques pures demeuraient l'algèbre, l'analyse et la géométrie. A la fin du XX^e siècle, c'étaient l'algèbre, l'analyse et la topologie* » STEWART Ian, *op. cit.*, p. 124.

- xlili Leonhard Euler, https://fr.wikipedia.org/wiki/Leonhard_Euler
- xliv Seven bridges of Königsberg, https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg
- xliv *Solutio problematis ad geometriam situs pertinentis* in : *Commentarii academiae scientiarum imperialis Petropolitanae* (CASP), volume 8 :1736, publié en 1741, pp. 128-140 (<http://eulerarchive.maa.org/publications/journals/Commentarii.html>). Texte original en latin disponible sur : <http://eulerarchive.maa.org/pages/E053.html> ; traductions en anglais : *The solution of a problem relating to the geometry of position* in : NEWMAN James R., *World of Mathematics*, 1956 and in : BIGGS Norman, LLOYD E. Keith, WILSON Robin J., *Graph Theory 1736-1936*, Clarendon Press, 1976.
- xlvi Traduction française en format pdf : COUPY Emile, *Solution d'un problème appartenant à la géométrie de situation, par Euler*, Nouvelles annales de mathématiques 1^{ère} série, tome 10, (1851), pp. 106-119 (http://www.numdam.org/article/NAM_1851_1_10_106_1.pdf).
- xlvi Kaliningrad, <https://fr.wikipedia.org/wiki/Kaliningrad>
- xlvi Graph theory, <https://www.britannica.com/topic/graph-theory>; *Histoire de la topologie*, <https://leseditionsdeschavannes.com/2016/05/01/histoire-de-la-topologie/> ; BARILIER Etienne, *op. cit.*, p. 42.
- xlvi Schéma tracé de la main de Leonhard Euler in : *Solutio problematis ad geometriam situs pertinentis*, in *Commentarii academiae scientiarum Petropolitanae* 8, 1741, pp. 128-140, disponible sur : <http://eulerarchive.maa.org/docs/originals/E053.pdf> (page 2) ; EULER Leonard in : COUPY E., *op. cit.*, figure 1, p. 110 ch. 6.
- i Les trois schémas originaux, plus simples et que j'ai complétés, proviennent de : *Problème des sept ponts de Königsberg*, https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg.
- ii 1646-1716, Gottfried Wilhelm Leibniz, https://fr.wikipedia.org/wiki/Gottfried_Wilhelm_Leibniz
- iii « Topology developed as a field of study out of geometry and set theory, through analysis of concepts such as space, dimension, and transformation. Such ideas go back to Gottfried Leibniz, who in the 17th century envisioned the geometria situs (Greek-Latin for "geometry of place") and analysis situs (Greek-Latin for "picking apart of place") » Topology, <https://en.wikipedia.org/wiki/Topology>
- lii « Outre cette partie de la géométrie qui traite des grandeurs et qui a été de tout temps cultivée avec beaucoup de zèle, il en est une autre, jusqu'à nos jours complètement inconnue, dont Leibnitz a fait le premier mention et qu'il appela géométrie de position. D'après lui, cette partie de la géométrie s'occupe de déterminer seulement la position et de chercher les propriétés qui résultent de cette position; dans ce travail, il n'est besoin, ni d'avoir égard aux grandeurs elles-mêmes, ni de les calculer ; mais il n'est pas encore assez bien établi quels sont les problèmes de ce genre appartenant à la géométrie de position, et quelle méthode il faut employer pour les résoudre » (EULER Leonhard in : COUPY E., *op. cit.*, p. 107 ch. 1).
- lii « Sans parler de théorie des graphes, Euler était tout de même conscient que son travail reposait sur des idées nouvelles. Il a affirmé que son article portait sur la géométrie de position, une nouvelle discipline mathématique mentionnée brièvement par Leibniz au siècle précédent mais largement inconnue » (DEGUIRE Paul, *Les Ponts de Königsberg*, Département de mathématiques et statistiques, Université de Moncton, Canada, p. 4 disponible sur : http://www.umoncton.ca/umcm-sciences-mathstat/files/umcm-sciences-mathstat/wf/wf/pdf/les_ponts_de_konigsberg_1.pdf
- lii EULER Leonhard in : COUPY E., *op. cit.*, p. 107 ch. 1 ; BARILIER Etienne, *op. cit.*, p. 42.
- lii EULER Leonhard in : COUPY E., *op. cit.*, p. 107 ch. 3.
- lii Seven bridges of Königsberg, https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg
- lii EULER Leonhard in : COUPY E., *op. cit.*, pp. 108-109 ch. 4 et 5.
- lii Expliquée en langage moderne, la démarche d'Euler donne cela : « One replaces each land mass with an abstract "vertex" or node, and each bridge with an abstract connection, an "edge", which only serves to record which pair of vertices (land masses) is connected by that bridge. The resulting mathematical structure is called a graph. Since only the connection information is relevant, the shape of pictorial representations of a graph may be distorted in any way, without changing the graph itself. Only the existence (or absence) of an edge between each pair of nodes is significant. For example, it does not matter whether the edges drawn are straight or curved, or whether one node is to the left or right of another » (Seven bridges of Königsberg, https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg).
- lii EULER Leonhard in : COUPY E., *op. cit.*, p. 111 ch. 9.
- lii « Quant à moi, j'ai fait de problème le suivant beaucoup plus général : Quelle que soit la figure du fleuve et sa distribution en bras, et quel que soit aussi le nombre des ponts, trouver si une personne peut traverser le fleuve en passant une seule fois sur chaque pont » (EULER Leonhard in : COUPY E., *op. cit.*, pp. 107-108 ch. 2.). Et de préciser : « C'est pourquoi j'ai inventé une règle par le secours de laquelle, tant pour cette question que pour

toutes celles du même genre, il est facile de discerner si un tel arrangement des lettres peut ou non avoir lieu » (EULER Leonhard in : COUPY E., *op. cit.*, p. 110 ch. 7.).

^{lxii} DEGUIRE Paul, *op. cit.*, p. 3 ; « Une telle promenade n'existe pas, et c'est Euler qui donna la solution de ce problème en caractérisant les graphes que l'on appelle aujourd'hui « eulériens » en référence à l'illustre mathématicien, à l'aide d'un théorème dont la démonstration rigoureuse ne fut en fait publiée qu'en 1873, par Carl Hierholzer » in *Problème des sept ponts de Königsberg*, https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg; Leonhard Euler, https://fr.wikipedia.org/wiki/Leonhard_Euler

^{lxiii} DEGUIRE Paul, *op. cit.*, p. 3 ; « Une telle promenade n'existe pas, et c'est Euler qui donna la solution de ce problème en caractérisant les graphes que l'on appelle aujourd'hui « eulériens » en référence à l'illustre mathématicien, à l'aide d'un théorème dont la démonstration rigoureuse ne fut en fait publiée qu'en 1873, par Carl Hierholzer » in *Problème des sept ponts de Königsberg*, https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg; Leonhard Euler, https://fr.wikipedia.org/wiki/Leonhard_Euler

^{lxiv} « Quel que soit donc le cas proposé, on pourra très facilement reconnaître sur-le-champ, au moyen de la règle suivante, si le passage une seule fois sur tous les ponts est ou non possible. S'il y a plus de deux régions auxquelles conduisent un nombre impair de ponts, vous pouvez affirmer avec certitude qu'un tel passage est impossible. Mais si l'on est seulement conduit à deux régions par un nombre impair de ponts, le passage est possible, mais en commençant sa course par l'une ou l'autre de ces deux régions. Enfin, s'il n'y a aucune région à laquelle on soit conduit par un nombre impair de ponts, alors le passage pourra avoir lieu, comme on le désire, et en commençant sa marche par telle région qu'on voudra. Cette règle satisfait pleinement au problème proposé » (EULER Leonhard in : COUPY E., *op. cit.*, p. 118 ch. 20.). Wikipédia précise : « Euler shows that the possibility of a walk through a graph, traversing each edge exactly once, depends on the degrees of the nodes. The degree of a node is the number of edges touching it. Euler's argument shows that a necessary condition for the walk of the desired form is that the graph be connected and have exactly zero or two nodes of odd degree. This condition turns out also to be sufficient—a result stated by Euler and later proved by Carl Hierholzer. Such a walk is now called an Eulerian path or Euler walk in his honor. Further, if there are nodes of odd degree, then any Eulerian path will start at one of them and end at the other. Since the graph corresponding to historical Königsberg has four nodes of odd degree, it cannot have an Eulerian path » (*Seven bridges of Königsberg*, https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg).

^{lxv} « The only important feature of a route is the sequence of bridges crossed. This allowed him to reformulate the problem in abstract terms (laying the foundations of graph theory), eliminating all features except the list of land masses and the bridges connecting them » (*Seven bridges of Königsberg*, https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg). La résolution du *Problème des sept ponts de Königsberg* est considérée comme l'acte fondateur de la topologie et de la théorie des graphes (BARILIER Etienne, *op. cit.*, pp. 41-42 ; *Seven bridges of Königsberg*, https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg; *Problème des sept ponts de Königsberg*, https://fr.wikipedia.org/wiki/Probl%C3%A8me_des_sept_ponts_de_K%C3%B6nigsberg; CARLSON Stefan C., *Königsberg bridge problem*, Encyclopaedia Britannica (<https://www.britannica.com/science/Konigsberg-bridge-problem>)).

^{lxvi} DEGUIRE Paul, *op. cit.*, p. 4.

^{lxvii} BARILIER Etienne, *op. cit.*, pp. 41-42.

^{lxviii} Geometria situs. « The term topology was introduced by Johann Benedict Listing in the 19th century, although it was not until the first decades of the 20th century that the idea of a topological space was developed. By the middle of the 20th century, topology had become a major branch of mathematics ». Topology, <https://en.wikipedia.org/wiki/Topology>

^{lxix} DEGUIRE Paul, *op. cit.*, p. 4 ; STEWART Ian, *op. cit.*, p. 123.

^{lxx} Traduction de : CARLSON Stefan C., *Graph Theory*, Encyclopaedia Britannica (<https://www.britannica.com/topic/graph-theory>). Pour d'autres exemples, cf. supra p. 3.

^{lxxi} Pour les références au texte d'Euler en différentes langues, cf. supra notes xlv et xlvii ; DEGUIRE Paul, *Les Ponts de Königsberg*, Département de mathématiques et statistiques, Université de Moncton, Canada disponible sur : http://www.umoncton.ca/umcm-sciences-mathstat/files/umcm-sciences-mathstat/wf/wf/pdf/les_ponts_de_konigsberg_1.pdf; ou alors de visionner cette courte vidéo de VAN DER VIEREN Dan, *How the Königsberg bridge problem changed mathematics*, Ted-Ed, 2016/09/01 sur : <https://archive.org/details/youtube-nZwSo4vfw6c>; ou de lire le passionnant ouvrage de BARILIER Etienne, *Leonhard Euler. La clarté de l'esprit*, Lausanne, Presse polytechniques et universitaires romandes, Savoir Suisse,

pp. 41ss ; et demeure bien évidemment l'article *Seven bridges of Königsberg*, de l'incontournable encyclopédie en ligne *Wikipedia* (https://en.wikipedia.org/wiki/Seven_Bridges_of_K%C3%B6nigsberg).

lxiii Pour des informations complémentaires sur ces développements ultérieurs voir : *Topologie*, <https://fr.wikipedia.org/wiki/Topologie>

lxiii Voir à ce sujet le Billet #4A - *Variété, le 3e V du Big Data*.

lxiv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 133.

lxv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 182.

lxvi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 182.

lxvii *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes

lxviii Source : *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes (Auteur : Goel — Travail personnel, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=48274506>)

lxix *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes

lxx MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 145.

lxxi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 149.

lxxii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 148.

lxxiii Billet #4A - *Variété, le 3e V du Big Data*, p. 5.

lxxiv MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 178.

lxxv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 177.

lxxvi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 179.

lxxvii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 149.

lxxviii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 178 et 188.

lxxix <https://www.sixdegreesofwikipedia.com/?source=Digitization&target=Justice>

xc Algorithmes fréquents. Pour plus de détails à ce sujet cf. MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 31, 218, 430, 438.

xc Billet #3D – *Hadoop- une évolution : un écosystème, le Big Data, et la résolution des conflits !*, pp. 14 et suivantes.

xcii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 150.

xciii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 179.

xciv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 179.

xcv MUELLER John Paul, MASSARON Luca, *op. cit.*, 206.

xcvi MUELLER John Paul, MASSARON Luca, *op. cit.*, 206.

xcvii MUELLER John Paul, MASSARON Luca, *op. cit.*, 206.

xcviii MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 179-180 ; *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes

xcix MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 150.

c MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 180.

ci MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 178 et 187.

cii Voir pp. 14 et suivantes.

ciii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 187.

civ MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 179.

cv *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes (Auteur : Goel — Travail personnel, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=48274506>); MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 179.

cvi Cf. supra p. 8.

cvi PENTLAND Alex, *Social Physics : How social networks can make us smarter*, New York, Penguin Books, p. 2.

cvi Réseau, <https://fr.wikipedia.org/wiki/R%C3%A9seau>

cix MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 178

cx *Théorie des graphes*, https://fr.wikipedia.org/wiki/Th%C3%A9orie_des_graphes

cx World Wide Web, https://fr.wikipedia.org/wiki/World_Wide_Web

cxi Pour des définitions, voir supra p. 9, *Directionnalité*.

cxi World Wide Web, https://fr.wikipedia.org/wiki/World_Wide_Web

cxi « Consider the directed graph whose nodes correspond to static pages on the web, and whose arcs correspond to links between these pages » BRODER Andrei and al., *op. cit.*, Introduction, p. 1.

cxi « We study various properties of this graph including its diameter, degree distributions, connected components, and macroscopic structure » BRODER Andrei and al., *op. cit.*, Introduction, p. 1.

cxi BRODER Andrei and al., *op. cit.*, Abstract, p. 1.

- cxvii CHRISTAKIS Nicholas, FOWLER James, *Connected : the amazing power of social networks and how they shape our lives*, London, Harper Press, Preface, p. xi.
- cxviii BABINET Gilles, *Big Data : penser l'homme et le monde autrement*, Paris, Le Passeur, 2015, p. 26 (format poche : <http://www.eyrolles.com/Informatique/Livre/big-data-penser-l-homme-et-le-monde-autrement-9782368904923>).
- cxix *Les topologies*, <https://openclassrooms.com/courses/les-reseaux-de-zero/les-topologies-1>
- cxx *Topologie de réseau*, https://fr.wikipedia.org/wiki/Topologie_de_r%C3%A9seau
- cxxi Les topologies, OpenClassrooms (<https://openclassrooms.com/courses/les-reseaux-de-zero/les-topologies-1>); https://fr.wikibooks.org/wiki/Les_r%C3%A9seaux_informatiques/Les_topologies_logiques
- cxixii Billet #3B - Hadoop – une architecture distribuée, un nouveau paradigme infrastructurel.
- cxixiii Billet #3C - Hadoop – des logiciels : MapReduce et Hadoop Distributed File System, un nouveau paradigme de programmation.
- cxixiv CHOKOGOUE Juvénal, *Hadoop : devenez opérationnel dans le monde du Big Data*, St-Herblain, ENI, 2017, pp. 90-91, 101 (<https://m.editions-eni.fr/livre/hadoop-devenez-operationnel-dans-le-monde-du-big-data-9782409007613#>).
- cxixv Pour les définitions de ces termes, cf. supra p. 9, *Directionnalité*.
- cxixvi CHOKOGOUE Juvénal, *op. cit.*, pp. 90-91, 101.
- cxixvii Modèle mathématique du WEB « Le World Wide Web, en tant qu'ensemble de ressources hypertextes, est modélisable en graphe orienté possédant des cycles avec les ressources pour sommets et les hyperliens pour arcs. Comme le graphe est orienté, certaines ressources peuvent constituer des puits, ou moins formellement des culs-de-sac : il n'existe aucun chemin vers le reste du web. À l'inverse, certaines ressources peuvent constituer des sources : il n'existe aucun chemin depuis le reste du web » (*World Wide Web*, https://fr.wikipedia.org/wiki/World_Wide_Web).
- cxixviii Voir Billet #3C - Hadoop – des logiciels : MapReduce et Hadoop Distributed File System, un nouveau paradigme de programmation, pp. 5 et suivantes.
- cxixix MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 243 et pp. 238-239.
- cxixx MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 243.
- cxixxi CARDON Dominique, *Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google*, Réseaux 2013/1 (n° 177), p. 65. Disponible sur : <https://www.cairn.info/revue-reseaux-2013-1-page-63.htm>
- cxixxii Pour plus de détails, voir FD ALGO pp. 239-240-241-242. Définition Techniques *white hat* et *black hat* sur : *optimisation pour les moteurs de recherche*, https://fr.wikipedia.org/wiki/Optimisation_pour_les_moteurs_de_recherche
- cxixxiii MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 240 à 242
- cxixxiv PAGE Lawrence, BRIN Sergey, MOTWANI Rajeev, WINOGRAD Terry, *The PageRank Citation Ranking: Bringing Order to the Web*, Introduction and Motivation, Technical Report. Stanford InfoLab. (29/01/1998 ; Disponible en format pdf sur : <http://ilpubs.stanford.edu:8090/422/>).
- cxixxv Google, <https://fr.wikipedia.org/wiki/Google>; FD ALGO p. 240; Pour plus de détails sur ce brevet intitulé *Method for Node Ranking in a Linked Database* (<https://patents.google.com/patent/US6285999B1/en>).
- cxixxvi BRIN Sergei, PAGE Lawrence, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia (<http://ilpubs.stanford.edu:8090/361/>).
- cxixxvii BRIN Sergei, PAGE Lawrence, *op. cit.*, Abstract, p. 1.
- cxixxviii « (...) is designed to crawl and index the Web efficiently and produce much more satisfying search results than existing systems » (BRIN Sergei, PAGE Lawrence, *op. cit.*, Abstract, p. 1).
- cxixxix Texte original : « (...) a large-scale search engine which makes heavy use of the structure present in hypertext » (BRIN Sergei, PAGE Lawrence, *op. cit.*, Abstract, p. 1).
- cxl Texte original : « there are new technical challenges involved with using the additional information present in hypertext to produce better search results » (BRIN Sergei, PAGE Lawrence, *op. cit.*, Abstract, p. 1).
- cxli Hyperlien, <https://fr.wikipedia.org/wiki/Hyperlien>
- cxlii CARDON Dominique, *op. cit.*, p. 71.
- cxliiii « Also we look at the problem of how to effectively deal with uncontrolled hypertext collections where anyone can publish anything they want » (BRIN Sergei, PAGE Lawrence, *op. cit.*, Abstract, p. 1).
- cxliiv *World Wide Web*, https://fr.wikipedia.org/wiki/World_Wide_Web
- cxliv GHERNAOUTI S., DUFOUR A., *Internet*, Paris, Presses universitaires de France, 12^e éd., 2017, p. 25.
- cxlvi *Hypertexte*, <https://fr.wikipedia.org/wiki/Hypertexte>
- cxlvii *World Wide Web*, https://fr.wikipedia.org/wiki/World_Wide_Web

-
- cxlviii *Hyperlien*, <https://fr.wikipedia.org/wiki/Hyperlien>
- cxlix *PageRank*, <https://fr.wikipedia.org/wiki/PageRank>. L'illustration ci-après provient d'une copie d'écran de la page *PageRank* sur Wikipedia (notamment le schéma « fantaisie » en haut à droite), que j'ai ensuite complétée.
- cl BABINET Gilles, *op. cit.*, p. 26.
- cli MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 237-238.
- clii « *In this paper, we take advantage of the link structure of the Web to produce a global «importance» ranking of every web page. This ranking, called Pagerank, helps search engines and users quickly make sense of the vast heterogeneity of the World Wide Web* » (PAGE Lawrence, BRIN Sergey, MOTWANI Rajeev, WINOGRAD Terry, *op. cit.*, p. 1 ch. 1).
- cliii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 241, voir également p. 244.
- cliv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 241.
- clv Texte original : « *PageRank can be thought of as a model of user behavior* » (BRIN Sergei, PAGE Lawrence, *op. cit.*, p. 4 ch. 2.1.2).
- clvi Texte original : « *The importance of a Web page is an inherently subjective matter, which depends on the readers interests, knowledge and attitudes. But there is still much that can be said objectively about the relative importance of Web pages. This paper describes PageRank, a method for rating Web pages objectively and mechanically, effectively measuring the human interest and attention devoted to them. We compare PageRank to an idealized random Web surfer* » (PAGE Lawrence, BRIN Sergey, MOTWANI Rajeev, WINOGRAD Terry, *op. cit.*, Abstract, p. 1).
- clvii *PageRank*, <https://fr.wikipedia.org/wiki/PageRank>; cf. supra p. 10 *Centralité*.
- clviii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 244.
- clix PAGE Lawrence, BRIN Sergey, MOTWANI Rajeev, WINOGRAD Terry, *The PageRank Citation Ranking: Bringing Order to the Web*, Introduction and Motivation, Technical Report. Stanford InfoLab. (29/01/1998 ; disponible sur : <http://ilpubs.stanford.edu:8090/422/>); BRIN Sergei, PAGE Lawrence, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*, Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia (<http://ilpubs.stanford.edu:8090/361/>).
- clx « *Un lien retour (aussi appelé lien entrant ou lien arrivant ; en anglais, inbound link, inlink ou backlink) est un hyperlien pointant vers un site ou une page Web* » (Lien retour, https://fr.wikipedia.org/wiki/Lien_retour).
- clxi PAGE Lawrence, BRIN Sergey, MOTWANI Rajeev, WINOGRAD Terry, *op. cit.*, p. 3, ch. 2.3 - Propagation Of Ranking Through Links.
- clxii *PageRank*, <https://en.wikipedia.org/wiki/PageRank> et le commentaire qui l'accompagne : « *Mathematical PageRanks for a simple network, expressed as percentages. (Google uses a logarithmic scale.) Page C has a higher PageRank than Page E, even though there are fewer links to C; the one link to C comes from an important page and hence is of high value. If web surfers who start on a random page have an 85% likelihood of choosing a random link from the page they are currently visiting, and a 15% likelihood of jumping to a page chosen at random from the entire web, they will reach Page E 8.1% of the time. (The 15% likelihood of jumping to an arbitrary page corresponds to a damping factor of 85%.) Without damping, all web surfers would eventually end up on Pages A, B, or C, and all other pages would have PageRank zero. In the presence of damping, Page A effectively links to all pages in the web, even though it has no outgoing links of its own* ».
- clxiii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 243.
- clxiv ABITEBOUL Serge, DOWEK Gilles, *Le temps des algorithmes*, Paris, Le Pommier, 2017, p. 95. (<https://www.editions-lepommier.fr/le-temps-des-algorithmes>).
- clxv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 253.
- clxvi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 252; DEAN Brian, *Google's 200 Ranking Factors: The Complete List* (2018), Backlinko, 26/05/2018 (<https://backlinko.com/google-ranking-factors>).
- clxvii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 253.
- clxviii *RankBrain*, <https://en.wikipedia.org/wiki/RankBrain>
- clxix *Google Hummingbird*, https://en.wikipedia.org/wiki/Google_Hummingbird
- clxx *Définitions de Rankbrain*, Définitions SEO (<https://www.definitions-seo.com/definition-de-rankbrain/>).
- clxxi *RankBrain : latest on Google algortithm update*, Infographic design team, 25/01/2018 (<https://www.infographicdesignteam.com/blog/rankbrain-latest-google-algorithm-update/>).
- clxxii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 254.
- clxxiii MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 253-254.
- clxxiv MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 244.
- clxxv Traduction de *PageRank*, <https://en.wikipedia.org/wiki/PageRank>
- clxxvi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 244.

-
- clxxvii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 253.
- clxxviii *More than just a Web Search algorithm: Google's PageRank in non-Internet contexts*, 03/11/2014 (<https://blogs.cornell.edu/info2040/2014/11/03/more-than-just-a-web-search-algorithm-googles-pagerank-in-non-internet-contexts/>).
- clxxix MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 181, 197 et 427.
- clxxx MUELLER John Paul, MASSARON Luca, *op. cit.*, pp. 218 et suivantes.
- clxxxi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 218.
- clxxxii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 187.
- clxxxiii MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 181.
- clxxxiv Pour ces 3 derniers éléments, cf. MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 253.
- clxxxv Traduction de *PageRank*, <https://en.wikipedia.org/wiki/PageRank>
- clxxxvi MUELLER John Paul, MASSARON Luca, *op. cit.*, p. 89, 182, 226
- clxxxvii <https://www.sixdegreesofwikipedia.com/?source=Digitization&target=Justice>; cf. supra p. 9.
- clxxxviii Voir à ce propos le Billet #3A – *Amélioration des moteurs de recherche et naissance d'Hadoop*, p. 2.
- clxxxix BABINET Gilles, *op. cit.*, p. 26.
- cxc Cf. Billet #3C *Hadoop – des logiciels : MapReduce et Hadoop Distributed File System, un nouveau paradigme de programmation*, pp. 3 et suivantes. Le parallélisme, pierre angulaire du passage de la centralisation vers la distribution, constitue un élément clé de l'avènement du Big Data.