Synthèse bibliographique
Fouille de réseaux sociaux en ligne

Matthieu,
Gaëtan,
Benoît,
Stéphane

Table des matières

1 Introduction **3**

2 Les processus pour la fouille de données **4**
2.1 Prédiction de liens 4
2.2 Identification des acteurs importants et des experts 7
2.3 Extraction de communautés 9

3 Vie privée dans les réseaux sociaux **14**
3.1 Lecture des données publiques 14
3.2 Lecture des données privées 15
3.3 Fouille de données dans le respect de vie privée 17

4 Marketing des données **20**
4.1 Prospection de nouveaux clients 20
4.2 Marketing viral 22
4.3 Ciblage des entités influentes 24

5 Conclusion **27**

A Modélisation des réseaux sociaux **28**
A.1 Propriétés des réseaux sociaux 30
A.1.1 Vocabulaire 30
A.1.2 Loi de distribution des degrés 31
A.1.3 Diamètre et « petit monde » 33
A.1.4 Dynamique et évolution 34
A.1.5 Clusterisation 35
A.2 Modélisation des réseaux sociaux 36
A.2.1 Modélisation de la structure d'un réseau social par génération aléatoire 36
A.2.2 Modélisation de la structure d'un réseau social par les « petits mondes » 37
A.2.3 Modélisation de la dynamique et de l'évolution 40

1 Introduction

Les réseaux sociaux sont omniprésents depuis l'avènement d'Internet.
Ils permettent aux différents utilisateurs d'interagir en communauté et de se regrouper selon des critères qui leur sont importants.

Ces réseaux sociaux sont de différents types. Certains sont connus de tous (Facebook, Twitter, LinkedIn) et comptent des millions de membres. D'autres exploitent des niches moins connus et peuvent passer relativement inaperçus ou rester confidentiels, tels les réseaux d'entreprise.
Enfin, certains des échanges peuvent aussi être assimilés à des réseaux sociaux : c'est le cas des mails et des SMS, qui définissent des échanges entre différents groupes d'individus.

Tous ces réseaux sociaux amassent de très nombreuses données : les amis, les messages, les images, la fréquence d'utilisation... tous ces échanges et informations sont soigneusement enregistrés. Dès lors se pose le problème de l'exploitation de cette masse d'informations.

Il faut tout d'abord modéliser le réseau sous forme mathématique. La structure de base est bien entendu le graphe : l'analyse des figures produites permet de tirer un grand nombre d'informations, et aussi de prédire en partie l'évolution future du réseau. Tous ces mécanismes seront abordés dans la partie 2 (page 4).

Dès lors se pose la question de la vie privée. Les données récoltés et analysées, comme dit précédemment, permettent d'intuiter un grand nombre d'informations qu'un utilisateur ne souhaite pas forcément divulguer. Au cœur des débats se trouve bien entendu l'éthique de la fouille : des mécanismes sont donc mis en palce pour protéger l'utilisateur de la curiosité parfois mal placée du fouilleur. Ces mécanismes sont détaillés dans la partie 3 (page 14).

Dans la partie 4 (page 20) seront détaillés les mécanismes mis en place pour rentabiliser ces informations en créant des stratégies marketing depuis les données précédemment analysées et protégées.

2 Les processus pour la fouille de données

Certains résultats issus de l'analyse des réseaux sociaux doivent être assimilés avant d'aborder le contenu de cette section. Ces résultats sont détaillés dans l'annexe A de ce document. On pourra rappeler ici quelques propriétés caractérisant les graphes représentants des réseaux sociaux :

- *les degrés des nœuds du graphe suivent une distribution en loi de puissance ;*
- *le diamètre de ces graphes est généralement faible et en* $O(\log n)$ *;*
- *le taux de clusterisation est élevé ;*
- *ces graphes deviennent de plus en plus denses et leur diamètre diminue avec le temps.*

L'état de l'art pour la Fouille de Données dans les réseaux sociaux se divise en plusieurs axes principaux :

- prédiction de liens (pour améliorer l'efficacité du marketing viral par exemple) ;
- identification des acteurs importants et des experts ;
- extraction de communautés ;
- identification des rôles joués par les individus en fonction de leurs liens (*link mining*) ;
- diffusion de l'information (diffusion des épidémies, contagion, etc.) ;
- recommandations et confiance ;
- recherche dans les réseaux sociaux (amélioration des algorithmes de routage, etc.) ;
- anonymisation et Vie Privée.

Seuls les trois premiers points énumérés ci-dessus seront traités dans cette partie. On pourra noter que le dernier point fera l'objet d'une partie dédiée dans cette synthèse.

2.1 Prédiction de liens

Le problème de la prédiction de liens a été vu, pour l'instant, comme un problème de reconstitution de liens manquants sur un graphe. Il s'agit donc d'un procédé qui fait intervenir de l'apprentissage.

Modèle de prédiction selon la distance

L'approche de [LNK04] se base le principe que les liens se forment selon la distance topologique entre les deux nœuds du lien. Il s'agit donc pour eux d'une caractéristique de la topologie du graphe. Pour démontrer cela, ils ont effectué une simulation sur le réseau ArXiv en utilisant plusieurs méthodes de calcul de rang en se basant sur les données connues entre 1994 et 1996 et en essayant de prédire les liens apparus entre 1997 et 1999.

graph distance	(negated) length of shortest path between x and y
common neighbors	$\lvert\Gamma(x) \cap \Gamma(y)\rvert$
Jaccard's coefficient	$\frac{\lvert\Gamma(x)\cap\Gamma(y)\rvert}{\lvert\Gamma(x)\cup\Gamma(y)\rvert}$
Adamic/Adar	$\sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\log\lvert\Gamma(z)\rvert}$
preferential attachment	$\lvert\Gamma(x)\rvert \cdot \lvert\Gamma(y)\rvert$
Katz$_\beta$	$\sum_{\ell=1}^{\infty} \beta^\ell \cdot \lvert\mathsf{paths}_{x,y}^{\langle\ell\rangle}\rvert$ where $\mathsf{paths}_{x,y}^{\langle\ell\rangle}$:= {paths of length exactly ℓ from x to y} weighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle}$:= number of collaborations between x, y. unweighted: $\mathsf{paths}_{x,y}^{\langle 1\rangle}$:= 1 iff x and y collaborate.
hitting time	$-H_{x,y}$
stationary-normed	$-H_{x,y} \cdot \pi_y$
commute time	$-(H_{x,y} + H_{y,x})$
stationary-normed	$-(H_{x,y} \cdot \pi_y + H_{y,x} \cdot \pi_x)$ where $H_{x,y}$:= expected time for random walk from x to reach y π_y := stationary distribution weight of y (proportion of time the random walk is at node y)
rooted PageRank$_\alpha$	stationary distribution weight of y under the following random walk: with probability α, jump to x. with probability $1-\alpha$, go to random neighbor of current node.
SimRank$_\gamma$	$\begin{cases} 1 & \text{if } x = y \\ \gamma \cdot \frac{\sum_{a\in\Gamma(x)}\sum_{b\in\Gamma(y)} \mathsf{score}(a,b)}{\lvert\Gamma(x)\rvert\cdot\lvert\Gamma(y)\rvert} & \text{otherwise} \end{cases}$

FIGURE 1 – [LNK04] Pour information, définition des différentes mesures ayant été comparées entre elles durant cette étude.

Les résultats obtenus ne sont pas satisfaisants (le meilleur d'entre eux est correct sur seulement 16 % de ses prédictions) mais ceux-ci permettent clairement de montrer que la topologie des graphes joue quand même un rôle sur les liens entre les nœuds.

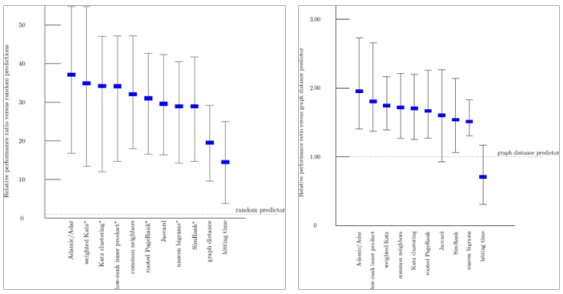

FIGURE 2 – [LNK04] Comparaison de la performance des mesures utilisées par rapport à une prédiction aléatoire et par rapport à une prédiction simplement basée sur la distance.

Modèle *Hierarchical Random Graphs*

Parmi les modèles proposés pour modéliser un réseau social a été récemment proposé *Hierarchical Random Graphs* par [CMN08]. Ce modèle se base sur le principe que les réseaux sont organisés hiérarchiquement. Cette approche permet entre autres de retrouver les propriétés caractéristiques des réseaux sociaux que sont les lois de distribution des degrés, des coefficients de clusterisation élevés et des petits chemins.

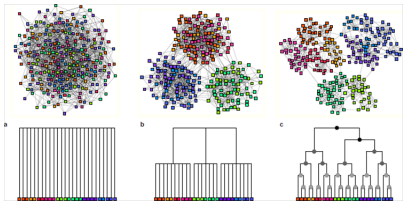

FIGURE 3 – [CMN08] À droite, un réseau dit hiérarchique avec son *Hierarchical Random Graph*. Le dendrogramme correspondant possède, pour chaque nœud, une probabilité p_r telle que les deux sous-arbres de ce nœud soient liés par un lien.

Cette approche n'a néanmoins pas fait beaucoup de bruit dans la communauté puisque celle-ci n'en n'est qu'à ses débuts. Cependant, des tests de prédiction de liens ont été effectués et se sont révélés concluants.

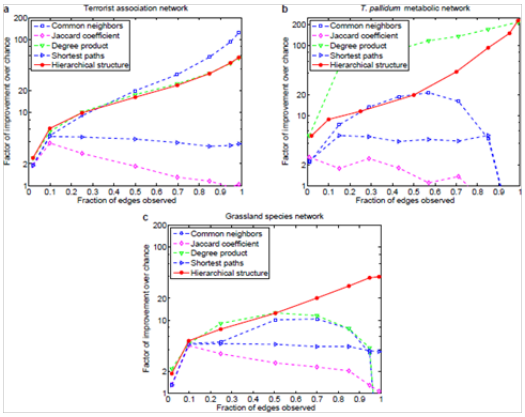

FIGURE 4 – [CMN08] Comparaisons des résultats des différentes mesures utilisées pour faire de la prédiction de liens. La méthode Hierarchical Random Graph est plutôt efficace en comparaison avec les autres.

2.2 Identification des acteurs importants et des experts

L'identification des acteurs et des experts au sein d'un réseau social est un problème issu du fait que nous disposions d'un nombre important de données sur les liens entre les différents nœuds d'un graphe. Cette connaissance sur les relations entre les individus a amené les chercheurs à étendre les tâches de Fouilles de Données classiques puisqu'il est envisageable de faire des classifications à partir des liens que possède un nœud (et non plus à partir de ses attributs seuls) par exemple.

Identification des acteurs importants

Cette approche de déduire certaines propriétés d'un nœud à partir de ses liens se retrouve dans la théorie de [Kle97], où il identifie les nœuds dit hubs ainsi que ceux qui sont dits autorités. L'algorithme le plus connu du grand public concernant cet aspect de la recherche est certainement l'algorithme utilisé par Google, *PageRank*, cité dans le document de [BP99].

Cette importance peut également être mesurée par diverses mesures basées sur la centralité des nœuds (centralité vis-à-vis des degrés, de la proximité, etc.).

Identification des experts

Une identification des experts sur un domaine est également possible et a fait l'objet de plusieurs recherches. Ces identifications sont basées sur une analyse d'informations issues de publications ou d'échanges informatisés (mails) pour identifier les différents acteurs d'un document et en déduire leur expertise sur des domaines particuliers.

Un exemple de recherches à ce sujet là est le travail de [STLS05] qui utilise les citations d'un document pour construire le profil *ExpertiseNet* d'un individu, constitué de plusieurs domaines dans lesquels l'individu en question est plus ou moins compétent. Cette compétence est ensuite quantifiée pour être restituée lors d'une requête.

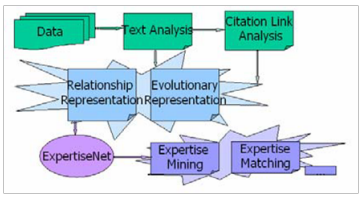

FIGURE 5 – [STLS05] Fonctionnement du système *ExpertiseNet*.

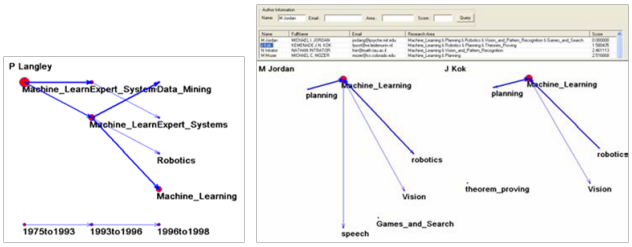

FIGURE 6 – [STLS05] Résultats obtenus après une requête sur *ExpertiseNet*. On notera que l'expertise d'un individu dans des domaines donnés est fonction du temps.

Un autre exemple de système d'identification d'experts est proposé dans les travaux de [LTZ+07], sur l'*Enterprise Oriented Search (EOS) Using Social Networks*, qui a amené à l'implémentation du moteur de recherche disponible à l'adresse suivante : http ://www.arnetminer.net, recensant des informations sur plus de 500 000 chercheurs dans le domaine de l'informatique et en bâtissant un réseau social basé sur leurs relations d'auteurs et de co-auteurs.

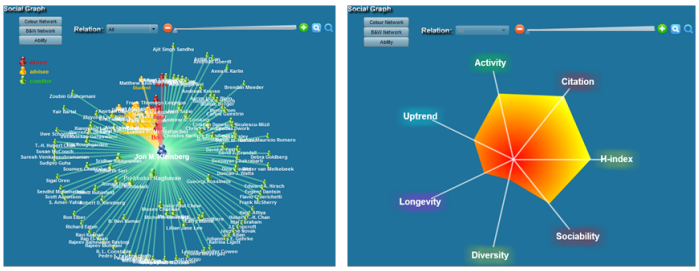

FIGURE 7 – Captures d'écran de l'outil ArnetMiner.

L'algorithme utilisé pour l'EOS est basé sur les mêmes principes que pour les autres algorithmes de recherche d'experts, avec l'utilisation de deux sources ici : la DBLP pour le profil d'un chercheur et le Citeseer pour les classements de publications et de citations.

2.3 Extraction de communautés

Il est possible d'utiliser un nombre important d'algorithmes classiques de détection de groupes (qui ne seront pas abordés dans ce document), basés sur les procédés suivants :

- partitionnement de graphe (par coupures minimales) [1] mais cela suppose de connaître à l'avance le nombre de partitions à réaliser ;
- clustering hiérarchique mais il s'agit d'une méthode en général coûteuse en $O(n^2 \log n)$ et qui n'est bien sûr pas adaptée aux structures n'ayant pas de hiérarchie à la base ;
- clustering en partitions (k-means) mais comme pour les procédés de partitionnement de graphe, le nombre de partitions est à définir à l'avance ;
- clustering spectral utilisant les représentations matricielles des graphes ainsi que leurs propriétés (valeurs propres) pour définir les clusters pour lequel nous pourrons nous reporter à [LV07].

Nous nous intéresserons ici à des méthodes adaptées à la topologie particulière des réseaux sociaux en lignes.

1. Pour rappel, un partitionnement d'un graphe est tel que chaque nœud appartient à un, et un seul, cluster.

Algorithme de Girvan-Newman

Une nouvelle approche de détection de communautés a été proposée par [GN01]. Cet algorithme est basé sur la décomposition d'un graphe en enlevant un par un les liens ponts, ces derniers étant identifiés par leur mesure de *betweeness*, à l'opposé des anciennes méthodes qui se contentaient de trouver les liens les plus forts.

En effet, les liens ponts sont les liens séparant deux communautés. Ceux-ci peuvent être identifiés à l'aide d'une mesure de betweeness qui quantifie le nombre de courts chemins dans l'ensemble du graphe passants par ce lien. L'algorithme est donc le suivant :

1. calculer les valeurs de betweeness pour tous les liens ;
2. enlever le lien qui a la plus grande valeur de betweeness (s'il y a égalité, en choisir un au hasard parmi les ex-æquo) ;
3. refaire le calcul des valeurs de betweeness pour tous les liens restants ;
4. répéter les étapes 2 et 3.

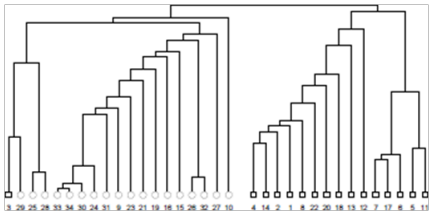

FIGURE 8 – [GN01] Application de l'algorithme de Givran-Newman sur le club de karaté de [Zac77]. Mis à part le nœud 3, la classification correspond aux observations.

Cet algorithme a une complexité de $O(m^2n)$ ou de $O(n^3)$ dans le cas d'un graphe peu dense avec m le nombre de liens et n le nombre de nœuds, un graphe étant peu dense si $m \ll n$, ce qui est bien le cas avec les réseaux sociaux.

Algorithme de Newman

L'algorithme proposé par [New03] permet de détecter des communautés dans des très grands graphes efficacement. Cet algorithme se base sur la mesure de *modularité* suivante : $Q = \sum_i (e_{ii} - a_i^2)$, avec e_{ii} la proportion

de liens internes à une communauté et $a_i = \sum_j (e_{ij})$ la proportion de liens faisant intervenir des nœuds de cette communauté avec des nœuds d'une autre communauté.

Cette mesure de modularité doit être maximisée en utilisant l'algorithme glouton suivant :

1. isoler chaque nœud dans des communautés séparées ;
2. calculer ΔQ pour chaque combinaison de communautés possibles ;
3. fusionner les communautés ayant le plus haut incrément ΔQ ;
4. répéter les étapes 2 et 3 jusqu'à avoir une valeur de la mesure Q maximale.

La complexité de cet algorithme est de $O((m+n) \times n)$ avec m le nombre de liens et n le nombre de nœuds ou de $O(n^2)$ dans le cas d'un graphe peu dense.

Amélioration de l'algorithme de Newman

Cet algorithme a été amélioré par [CNM04] pour atteindre une complexité de $O(md \log n)$ avec m le nombre de liens, d la profondeur du dendrogramme décrivant les communautés et n le nombre de nœuds. En sachant de plus qu'en général, $m \sim n$ et que $d \sim \log n$, on obtient une complexité de $O(n \log^2 n)$. Cette amélioration consiste juste en une meilleure gestion des données (par exemple, pas besoin de calculer ΔQ pour des communautés ne pouvant être fusionnées parce qu'il n'existe aucun pont ou meilleur suivi des ΔQ en utilisant un tas max).

Un jeu de test a été constitué à partir des données d'Amazon d'août 2003, soit un graphe de 409 687 nœuds et de 2 464 630 liens. Ce graphe représente des produits vendus sur Amazon, liés entre eux par des liens, un lien existant entre A et B si B est souvent acheté par les acheteurs de A. Le temps de calcul n'est pas donné mais voici la division en communautés qui a été obtenue :

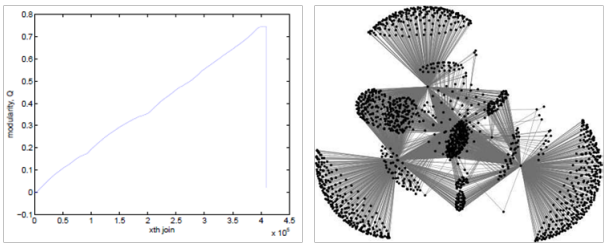

FIGURE 9 – [CNM04] Résultats de l'étude avec à gauche la modularité Q que l'on fait augmenter et avec à droite la représentation finale du graphe.

Détection de communautés superposées

L'ensemble des méthodes abordées jusqu'ici ne traitaient que de communautés ne pouvant être superposées. En réalité, de nombreux nœuds peuvent appartenir à plusieurs communautés en même temps. Cette problématique a été abordée dans les travaux de [PDFV05] et a débouché sur la méthode de percolation des cliques [2].

L'idée de cet algorithme est, à partir de k-cliques (des cliques avec k nœuds), de construire petit à petit les communautés, considérées ici comme un ensemble de cliques. Cette union est alors définie comme étant une « communauté k-clique ». Nous pouvons préciser que l'union de deux k-cliques est réalisée si ces cliques sont dites adjacentes donc si elles partagent $k-1$ nœuds.

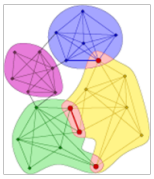

FIGURE 10 – [PDFV05] Superposition de communautés.

Détection de communautés dynamiques

Étant donné la non fiabilité actuelle et la marge d'amélioration possible des algorithmes de détection de communautés statiques, très peu de travaux sont entrepris dans le domaine de la détection des communautés évoluant

2. Pour rappel, une clique est un sous-graphe dont la particularité est qu'il existe un lien entre n'importe lequel des nœuds de ce sous-graphe.

dans le temps. Cette dynamique des communautés peut notamment s'exprimer par les phénomènes de croissance, de réduction, de fusion, de division, de naissance ou de mort de communautés.

3 Vie privée dans les réseaux sociaux

Avec l'émergence des réseaux sociaux, de nouvelles problématiques propres à la vie privée sont soulevées. Nous verrons dans cette partie que des dangers, souvent négligés, existent pour les internautes. Par ailleurs, nous nous intéresserons aux procédés pour exploiter les données de ces utilisateurs sans leur porter atteinte.

3.1 Lecture des données publiques

Dans cette première partie, nous verrons ensemble qu'il est relativement facile de récupérer légalement des informations personnelles.

Partage canapé et vie privée

Des chercheurs se sont penchés sur le site CouchSurfing [3] dans le but de déterminer les informations que les internautes considéraient confidentielles [PS09]. Sur ce site, les utilisateurs proposent littéralement leur canapé à qui voudrait séjourner temporairement chez eux, et ce, gratuitement. Pour entrer dans le réseau, l'internaute est appelé à partager un grand nombre d'informations avec les autres membres. Les résultats de cette étude soulèvent un certain nombre de points intéressants. Les trois quarts des utilisateurs affichent clairement leur code postal, genre, âge mais aussi leur photo de profil, leur emploi et leur nom complet. Pour certains, cela va plus loin avec la date de naissance, e-mail, photos de leur maison, adresse complète ou encore numéro de téléphone. Les raisons qui justifient de telles divulgations sont multiples. Aucun sondé ne semble particulièrement s'inquiéter par ces agissements. Certains évoquent le fait que ces mêmes informations sont accessibles dans l'annuaire téléphonique ou sur Google directement. Beaucoup reconnaissent divulguer des informations privées, mais ne voient pas l'intérêt qu'une personne pourrait avoir à les tracer et se doutent encore moins des exploitations malsaines, telles que l'usurpation d'identité. Mais le risque est bien là. Croiser ces informations avec d'autres sources peut permettre d'obtenir un profil très complet de la personne recherchée.

3. http ://www.couchsurfing.org/

Portrait volé

À ce sujet, le magazine français *Le Tigre* possède une rubrique intitulée Portrait Google. Le principe est très simple : dresser le profil d'une personne inconnue via les informations qu'elle laisse sur internet. Fin 2008, le journaliste Raphaël Metz s'est lancé dans le cadre de cette rubrique à l'assaut d'un certain Marc L., une personne choisie au hasard [Met09]. Après quelques recherches plus ou moins fructueuses sur les réseaux sociaux tels que Copains d'avant, Facebook, Flickr, Youtube, ou même Google, de nombreuses informations ont pu être récupérées. Tout cela lui a permis de rédiger une biographie relativement fournie en s'attachant à la fois aux relations amoureuses, expériences professionnelles, récits de vacances, ou encore coordonnées. Individuellement, comme dans l'étude précédente, les informations n'ont pas forcément une grande valeur. Mais une fois recoupées entre elles et en multipliant les sources, le résultat peut faire peur. En soit, la démarche n'a rien de scandaleuse, ces informations étant publiques. D'ailleurs, les recruteurs ont régulièrement recours à ces pratiques pour en savoir plus sur les personnes qu'ils s'apprêtent à engager. Le journaliste aurait même pu aller plus loin en s'inventant un personnage fictif cohérent avec le passé de sa cible, avec les mêmes intérêts ou en évoquant une rencontre passée imaginaire entre les deux dans un lieu identifié. Cet article a créé la polémique dans les médias [Lau09]. En découvrant ce récit de sa vie, l'intéressé a fait supprimer toutes les informations le concernant sur internet et a demandé d'anonymiser l'article en question.

Ainsi, les internautes laissent globalement un grand nombre d'informations personnelles sur les réseaux sociaux, et internet plus généralement, sans se soucier trop des conséquences sur leur vie privée.

3.2 Lecture des données privées

Nous avons pu voir précédemment que beaucoup d'informations à caractère privé étaient visibles et donc accessibles à tous sur les réseaux sociaux. Maintenant, nous allons nous intéresser à la lecture et surtout l'exploitation des données privées.

Google et la publicité

En 1998, Google a lancé son célèbre moteur. Quelques années plus tard, ce même moteur de recherche se retrouve en position de monopole. Cette position a permis à la société de Montain View d'être au premier rang des attentes des utilisateurs, en voyant les requêtes que ces derniers formulaient. Il n'a pas fallu attendre très longtemps pour que ces informations soient exploitées. En effet, dès l'année 2000, soit seulement deux ans après la création de l'entreprise, Google lance AdWords. Ce système va permettre d'afficher de la publicité en rapport avec les recherches des internautes. Ce nouveau type de publicité présente un gros intérêt car elle se situe à l'endroit et au moment où l'utilisateur a le plus de chance de s'y intéresser. C'est le premier pas de la société de ce que l'on appelle maintenant la publicité ciblée. Depuis ce temps, cette activité représente la majeure partie des revenus du groupe. En 2010, la publicité a atteint 96% des 29 milliards de ses entrées d'argent [Goo11].

Ciblage indiscret

Pour proposer une publicité de plus en plus ciblée, Google et les régies publicitaires ont besoin d'en savoir plus sur les internautes. Les cookies du navigateur internet sont d'ailleurs assez bavards à ce sujet. Mais cette recherche d'information sur les internautes est également le fil conducteur des services *gratuits* que Google lance. On peut notamment noter la suite bureautique [4], le partage de photos [5] ou encore le courrier électronique [6]. Ces services encouragent les internautes à déposer leurs informations personnelles directement sur les serveurs de Google. Ensuite, ce dernier peut analyser à souhait les données qu'il récupère, dans le but de proposer des publicités personnalisées. Ceci est particulièrement visible avec le service GMail où le courrier est automatiquement analysé par des *robots* [Dum04]. Par exemple, si un ami vous propose de l'accompagner à un séjour à la montagne, des offres pour des stations de ski apparaîtront dans la même fenêtre. Cette publicité est d'ailleurs très mal acceptée [Ast04]. Le courrier, qu'il soit électronique ou papier, a toujours été perçu comme une ressource critique au niveau de la confidentialité. Le fait qu'une entreprise privée le lise ouvertement est perturbant

4. Google Docs, https ://docs.google.com/
5. Picasa, http ://picasa.google.fr/
6. GMail, https ://mail.google.com/

vis-à-vis des libertés individuelles. Pour compléter ces informations sur les internautes, Google doit se tourner vers une nouvelle source très riche : les réseaux sociaux.

Réseaux sociaux : mine d'or du ciblage

Leader incontesté des réseaux sociaux, Facebook compte aujourd'hui plus de 800 millions d'utilisateurs [Fac11]. Très peu aurait imaginé qu'un jour plus d'un dixième de la population se soit créé un profil riche en informations personnelles sur un site internet privé. La monétisation de ses informations peut être très juteuse pour la jeune entreprise. Et pour cause ! Facebook propose aux annonceurs des options de ciblage extrême, comme les données démographiques, géographiques, ou encore les centres d'intérêt pour ne citer qu'eux. Il devient alors possible d'afficher des bannières aux personnes de la région parisienne, âgées de 26 à 30 ans et travaillant dans une entreprise concurrente. Quoi de mieux pour débaucher toute une équipe ? Ceci n'est qu'une des nombreuses possibilités envisageables. Cette activité est d'autant plus intéressante pour Facebook que les informations personnelles sont fournies gratuitement et volontairement par les utilisateurs. C'est justement à ce niveau que Google a pris du retard [Pui11]. Ce dernier a donc lancé l'été dernier son propre réseau social pour essayer de compléter son offre publicitaire. Mais il faut encore que le nombre d'utilisateurs atteigne une masse critique pour que la plate forme soit adoptée et fréquentée.

Ainsi, nous avons vu que les services gratuits sur internet étaient une mine d'or pour les régies publicitaires. Les réseaux sociaux sont particulièrement intéressants car ils fournissent *gratuitement* des informations pour un ciblage extrême et donc payant.

3.3 Fouille de données dans le respect de vie privée

Nous avons pu constater que la vie privée des individus pouvait être mis à mal sur internet, notamment à partir des réseaux sociaux. Pour essayer de préserver cette vie privée, différentes procédés ont été mis en place.

Anonymisation délicate

Simplement, nous pouvons penser qu'enlever les attributs tels que le nom et les coordonnées peut être suffisant mais ce n'est pas le cas. Comme nous

l'avons vu précédemment, les données peuvent être recoupées et ainsi mettre en péril tout la vie privée d'une personne. Une femme de 62 ans a justement été identifiée et ses habitudes mises à nue suite à une erreur sur une base de données d'AOL, bien que son nom et autres informations critiques aient été masqués [Bar06]. Nous pouvons également citer le cas du Netflix Prize Dataset. L'entreprise Netflix propose des services de location de films, films que ses clients peuvent noter. Pour améliorer leur service, la société a lancé un concours en mettant comme matière les notes enregistrées par presque un demi million de ses utilisateurs. Les informations d'identification avait préalablement été retirées, mais des chercheurs ont prouvé que cela n'était pas suffisant pour anonymyser l'ensemble [NS06]. En effet, ils ont développé pour l'occasion une méthode pour *désanonymiser* les données de ce type de jeu de données [7]. Ils ont ainsi démontré que la connaissance de quelques informations sur un individu permettait de savoir s'il était présent dans l'échantillon et de le retrouver. Alors, il devient possible d'en déduire des informations telles que ses orientations politiques et sexuelles, à partir des notations de films qu'il a effectuées. Nous allons donc nous pencher sur les méthodes reconnues pour travailler sur des données en préservant la vie privée.

Ajout de bruit

Une première approche consiste à faire varier de manière aléatoire les données recueillies [AA01]. Par exemple, dans le cas où l'étude cherche à analyser la distribution de l'âge des individus pour en établir une classification, il est possible d'ajouter une valeur de distorsion [8] à chacune des valeurs lors de l'acquisition des données. L'algorithme d'espérance-maximisation [Dem97] est recommandé pour assurer la vraisemblance de l'ensemble collecté. Généralement, la fouille de données n'a pas vraiment besoin d'informations individuelles mais plus de distributions sur les propriétés étudiées. Pour cette raison, il est tout à fait envisageable de rajouter du bruit aléatoire et contrôlé sur les données tout en retrouvant la distribution initiale du groupe et donc en préservant l'objet de l'étude. Cependant, cette approche n'est pas suffisante car elle permet de déduire certaines propriétés spécifiques du groupe d'étude. Il a été en effet démontré que l'ajout d'une petite perturbation aux données entraînait toujours une violation importante de la vie privée [DN04].

7. Très grande base avec des *micro-data* telles que les préférences sur un site web, les recommandations ou encore les transactions enregistrées

8. Via une distribution uniforme sur un segment ou une distribution Gaussienne

Randomisation

Pour remédier à ce problème, une approche a été entreprise en 2003 par des chercheurs de l'Université de Cornell et d'IBM pour compléter celle que nous venons d'évoquer [AES03]. Celle-ci consiste à appliquer une *amplification* aux valeurs. Cette amplification consiste à utiliser un opérateur de *randomisation* qui vérifie certaines conditions. Cet opérateur diminue la probabilité de retrouver la valeur initiale à partir de la valeur initial et inversement. L'intérêt de cette approche est qu'il n'est pas nécessaire d'avoir des connaissances ou de faire des hypothèses sur la distribution initiale. Par contre, il est difficile de quantifier son impact sur la précision de la fouille et donc de calibrer le paramétrage nécessaire à un certain degré respect de la vie privée. Il s'agit d'un domaine très actif de la recherche, notamment à cause de l'effervescence des réseaux sociaux, les approches présentées ne sont donc pas exhaustives et d'autres sont encore en cours de développement.

Nous avons pu constater que des moyens plus ou moins efficaces existaient pour anonymiser les sources pour la fouille de données. Cela pourrait, à l'avenir, réduire le risque de divulgations d'informations sur les internautes. Encore faut-il que les entreprises prennent conscience des dangers de leurs activités et qu'elles fassent les efforts nécessaires pour y remédier.

Afin de conclure cette partie, les réseaux sociaux peuvent révéler des informations précieuses sur leurs utilisateurs. L'apprentissage sur ces derniers peut être d'autant plus grand que les sources sont multiples et croisées. De plus, les groupes qui récupèrent ces données ne se gênent pas pour les monétiser, sans forcément se soucier des risques de divulgation. Des méthodes ont été développées pour justement essayer d'anonymiser ces données et ainsi limiter l'atteinte qu'une divulgation peut avoir sur les utilisateurs. Nous pouvons donc espérer que les risques diminuent à l'avenir.

4 Marketing des données

La fouille de données des réseaux sociaux représente une mine d'or d'informations qui fait l'objet d'un véritable marketing. Dans cette partie nous allons nous pencher sur l'utilisation de ces informations par des entreprises dans le but de prospecter de nouveaux clients ou de cibler les entités influentes afin d'introduire une stratégie marketing dite "virale".

4.1 Prospection de nouveaux clients

Une des principales applications du datamining est d'aider les entreprises à déterminer qui est susceptible de représenter un client potentiel, et qui n'est pas enclin à acheter son produit. Cette fouille de données peut jouer un rôle crucial dans la stratégie commerciale et marketing d'une entreprise. Il existe deux principaux types de démarchages qui sont le **marketing de masse** et le **marketing direct** [Bou05].
Le premier vise un public large appartenant à un même segment [9] via des médias de masse comme la télévision, le radio ou les journaux. C'est une méthode très efficace quand un produit est très demandé par les clients, comme par exemple l'électroménager après la seconde guerre mondiale. Aujourd'hui avec l'abondance de produits et la forte concurrence, les mœurs ont changé et ce type de marketing est jugé peu efficace.
Le second type de marketing, le marketing direct, cible un public à fort potentiel d'achat sur un produit spécifique. Ce public est visé en se basant seulement sur ses caractéristiques et ses besoins. La fouille de données joue un rôle clé dans le marketing direct puisqu'il permet de donner un modèle prévisionnel du comportement d'un client en fonction de ses derniers achats (habitudes de consommation, mode de paiement...) et d'informations personnelles (situation géographique, centres d'intérêts...).

Limiter ainsi sa cible à des clients ayant un fort potentiel d'achat est ce que l'on appelle en marketing le "scoring client". La prospection coûte cher aux entreprises, c'est pourquoi il leur est primordial de limiter le nombre de cibles à démarcher et de viser seulement les individus jugés les plus susceptibles d'acheter leurs produits. Avec l'arrivée d'Internet et des réseaux sociaux, les informations personnelles sur les individus sont beaucoup plus riches et facilement accessibles. Il est possible d'extrapoler ces informations et d'en

9. Un segment marketing (ou de clientèle) est une catégorie identifiable de clients visés.

déduire le comportement et les habitudes de consommation d'un individu. Ces informations une fois recoupées et analysées vont permettre d'assimiler ses individus à une niche de clients. Chaque niche est susceptible de présenter un fort potentiel d'achat à l'égard de certains produits et d'être réticent à d'autres.

Ainsi le datamining s'avère très efficace et donc très utilisé dans le marketing direct pour cibler un public à même d'être intéressé par le produit publicisé. Cependant, le datamining ne doit pas être considéré comme une solution miracle à l'ensemble des problèmes des entreprises. Il correspond à une avancée technologique qui permet de faire face au volume croissant des données collectées mais les entreprises devront instaurer un climat de confiance afin de ne pas porter atteinte à la vie privée des clients en exploitant les données collectées.

Les premières applications de datamining concernaient l'étude des tickets de caisse des clients de grande surface. Cela a permis de montrer que pour certaines catégories de clients les promotions mises en place pour des produits qu'ils avaient l'habitude d'acheter simultanément n'étaient pas efficaces et n'engendraient pas d'augmentation de chiffre d'affaires.

D'après les études de l'université de Western Ontario [LL98], la démarche pour cibler un public présentant un fort potentiel d'achat pour un produit spécifique est toujours similaire à celle-ci :

1. Récupérer les données de tous les clients possibles (retours des dernières prospections, informations personnelles sur des individus...)
2. Dataminer à cette base de données, c'est-à-dire :
   - ajouter des informations géo-démographiques
   - traiter les valeurs manquantes (déduites des autres données complètes)
   - séparer la base de données en deux jeux, qu'on appelle le "testing set" et "training set"
   - appliquer les algorithmes au "training set" pour estimer leur comportement en fonction des analyses déduites de l'étude du "testing set"
3. Évaluer le modèle déduit du "testing set". Recommencer la deuxième étape s'il n'est pas satisfaisant.
4. Utiliser le modèle déduit pour séparer les clients potentiels des non-acheteurs.
5. Promouvoir le produit aux clients potentiels retenus.

Cette démarche théorique et mathématique a fait ses preuves dans la pratique. Elle permet réellement de rentabiliser les opérations de prospection, en diminuant les coûts engendrés par l'acquisition de nouveaux clients. En mettant en place une solution de datamining, les entreprises essaient d'allonger la "durée de vie" d'un client en repérant les raisons et les risques de son départ. Par exemple, les sociétés de vente par correspondance réalisent des catalogues spécialisés en plus de leur catalogue généraliste. L'utilisation du datamining permet de sélectionner parmi les clients principaux ceux pour lesquels il est utile d'envoyer un catalogue spécialisé. C'est en effet grâce à l'historique des achats qu'il est possible de déterminer quel client est susceptible d'acheter un article sur catalogue spécialisé [PK09].

Cependant l'une des principales limites à cette démarche théorique vient du fait que la décision d'achat d'un client est considérée comme indépendante des autres alors qu'en réalité, cette décision est grandement influencée par ses collègues, ses amis, son entourage... Négliger l'influence d'un membre du réseau sur un autre est donc une mauvaise estimation du comportement d'un réseau puisque c'en est presque l'un des fondements. Cette influence est d'ailleurs ajourd'hui utilisée à des fins commerciales dans des campagnes marketing dites "virales".

4.2 Marketing viral

Comme nous venons de le voir, ignorer les effets de réseaux quand on décide quels clients démarcher peut mener à des décisions peu optimales. Le département informatique de l'université de Washington a étudié ce sujet et propose la modélisation d'une **valeur de réseau** [DR01], qui évalue l'influence qu'un client a sur les autres. En plus de la valeur intrinsèque du profit potentiel qu'un client peut représenter, cette étude propose donc d'ajouter une valeur qui représente l'importance de son réseau et donc son influence sur les autres clients. Ainsi un client potentiel représentant peu de profit mais étant très influent mériterait tout de même d'être démarché. Inversement, il serait redondant de démarcher un client avec un profit potentiel élevé puisqu'il a de fortes chances d'être influencé par son réseau. C'est le fondement de ce qui est appelé le **marketing viral**. Quand cette approche fonctionne, cela permet à l'entreprise d'augmenter significativement ses profits et de limiter son budget de communication et de publicité.

Le marketing viral est basé sur le réseau du "bouche à oreille" et propose un bien meilleur rapport qualité/prix qu'un marketing conventionnel puis-

qu'il s'appuie sur les clients eux-mêmes pour faire la promotion d'un produit. C'est là que réside l'un des éléments essentiels du marketing viral : chaque client devient un vendeur involontaire, simplement en utilisant le produit. Il est plus puissant que la publicité classique car il véhicule l'approbation implicite d'un ami. Un élément clé de la consommation de marque est d'ailleurs l'affiliation à un groupe : "je veux être un membre du groupe, le groupe étant dans ce cas les amis qui utilisent ledit produit". Ce type de marketing est essentiellement opéré via Internet, notamment sur les réseaux sociaux, et non les autres médias plus classiques auxquels les individus deviennent de plus en plus réticents. Le marketing viral exploite donc ces réseaux sociaux pour encourager les clients à partager des informations sur leurs produits. Ce type de marketing n'est donc pas efficace sur tous les types produits et il est difficile de mesurer son impact.

Un exemple des plus éloquents, développé dans le magazine Red Herring [Juv00], est celui du service de courriers électroniques "Free Hotmail" qui est passé de zéro à douze millions d'utilisateurs en moins d'un an et demi avec un budget publicitaire minime (cinquante mille dollars [10]), et ce grâce aux messages promotionnels inclus dans le pied de page de chaque email envoyé depuis leur service.

L'université de Canergie Mellon propose un modèle pour identifier quels types de produits sont susceptibles d'être promus par le marketing viral [JLA07]. Ce modèle est basé sur quelques principes évidents mais dont l'étude a levé quelques points intéressants.

Par analogie avec les maladies "virales", on peut distinguer les individus dits "sains", qui en l'occurence ne sont pas intéressés par le produit plubicisé ou qui n'en connaissent encore pas l'existence, des individus dits "infectés", qui utilisent le produit et qui sont donc suscpetibles d'influencer leur entourage à l'acheter également. Il apparaît notamment que comme toute épidémie, n'importe quelle interaction entre une personne "infectée" et une personne "saine" peut résulter à une contamination. Cependant, il semble que la probabilité de propagation diminue à interactions répétées. Par ailleurs, contrairement à une épidémie virale, le nombre de contacts entre une personne "infectée" avec une personne "saine" augmente la probabilité de contamination jusqu'à un certain seuil puis semble stagner. Ainsi, les in-

10. À la même époque, la compagnie Juno a dépensé plus de vingt millions de dollars sur une campagne conventionnelle pour un effet moindre.

dividus semblent être imperméables aux recommandations de leur entourage si le produit ne les intéresse vraiment pas.

Comme évoqué précédemment, chaque individu possède une valeur de réseau qui indique l'influence de celui-ci sur les autres clients potentiels. Certains semblent très influents, ils sont appelés les **super-spreaders** [JLA07] car ils "infectent" un grand nombre de personnes. Le modèle prend donc en compte qu'un super-spreader possède une plus forte probabilité de contamination. Néanmoins, il apparaît que lorsqu'un individu émet de plus en plus de recommandations, celles-ci s'avèrent de moins en moins efficaces. Ce qui laisserait supposer qu'un super-spreader a beaucoup d'influence sur quelques membres de son entourage, mais pas sur toutes les personnes qu'il connaît. Enfin, il semble que les produits les plus susceptibles de bénéficier d'une stratégie de marketing viral sont les produits technologiques ou idéologiques (religieux par exemple), puisqu'ils sont utilisés dans des contextes favorables comme une école, au travail ou sur un lieu de culte, des endroits ou tout un réseau est regroupé dans une même zone géographique et partage des centres d'intérêt communs.

Le marketing viral peut donc être très bénéfique puisque s'il fonctionne, le produit se répandra avec succès à moindre coup. Cependant c'est une stratégie risquée et limitée qui n'est pas adaptée à tous les produits et qui reste en grande partie abstraite malgré les nombreuses études sur le sujet.

4.3 Ciblage des entités influentes

La principale difficulté d'une stratégie de marketing virale est de cibler les individus ou entités les plus influents. La fouille de données semble être une méthode efficace pour déterminer qui est influent et qui possède un grand réseau. C'est en tout cas ce qu'indiquent les études de l'université de Cornell [DKT03].

Pour estimer la valeur du réseau d'un individu, une entreprise a besoin de connaître les relations entre les différents membres de son réseau. La source principale de ces informations est disponible sur Internet via les forums ou les réseaux sociaux. Les sites communautaires sont très souvent orientés sur les produits de consommation, il est donc possible de recueillir beaucoup d'informations sur les habitudes de consommation et les liens entre les différents membres de la communauté. On peut notamment trouver sur ces sites des tests complets de produits divers et variés, des évaluations, des classements, des comparaisons, des avis, des retours d'utilisation...

Supposons que nous disposons de données sur un réseau social, avec des estimations de l'influence mutuelle des individus, et que nous aimerions promouvoir un nouveau produit, qui nous l'espérons, sera adopté par une large partie de ce réseau.
Nous l'avons vu, le principe du marketing viral est qu'en ciblant d'abord quelques individus "influents" (par exemple, en leur donnant des échantillons gratuits des produits), il est possible de déclencher une cascade d'influence. Des amis recommandent le produit à d'autres amis, et ainsi de nombreux individus finiront par essayer. Mais comment doit-on choisir ces quelques individus clés (super-spreaders) à utiliser pour lancer ce processus au mieux ?

Une étude s'est fondée sur la fusion de deux algorithmes mathématiques imitant le comportement d'un réseau [DR02]. Le premier est basique et linéaire, le second prend en compte l'influence d'un membre sur les autres, il est dit "en cascade". Avec cette modélisation, il est possible de mesurer la dynamique d'un réseau de façon cyclique. À chaque cycle, un individu entre en contact avec son entourage et le contamine ou non.
Le problème réside alors dans le choix du noyau cible pour infecter le réseau le plus efficacement possible. Des chercheurs ont réfléchi à une fonction permettant de calculer pour chaque individu du réseau, en s'appuyant sur sa valeur de réseau, l'impact qu'il aura sur son entourage s'il est infecté [Dom04]. Cette fonction est ensuite optimisée afin de sélectionner un nombre défini de super-spreaders et en limitant les effets de redondance (par exemple en tentant de contaminer un individu déjà infecté). Ainsi est déterminé un noyau influent au sein d'un réseau, qui influence la grande majorité du réseau en quelques cycles seulement.

La limite à ses recherches est évidemment la qualité et la quantité des informations sur les relations entre les membres d'un réseau ainsi que l'évaluation de leur valeur de réseau. Cette technique de marketing est pour autant très prometteuse et continue à faire l'objet de nombreuses études.

La fouille de données s'avère donc être un outil puissant et très utilisé dans les stratégies de marketing actuelles. Appliquée à la mine d'or d'informations que représente aujourd'hui les réseaux sociaux et internet, cette science est largement exploitée par les entreprises. Celle-ci leur permet de cibler leur clientèle potentielle et de limiter leurs dépenses en prospection, en ne s'adressant qu'aux individus jugés susceptibles d'être demandeurs. C'est le fondement d'un nouveau type de marketing : le marketing viral. Il vise à cibler les personnes les plus influentes au sein d'un réseau afin de séduire un maximum d'individus en n'en démarchant qu'une petite partie. Ce marke-

ting très prometteur fait l'objet de nombreuses études et semble se peaufiner, notamment dans la sélection du noyau d'individus influents.

5 Conclusion

La fouille de données s'avère être un outil incroyablement riche et puissant lorsqu'elle est appliquée aux réseaux sociaux.

La masse de données produites est riche et remplie d'informations pertinentes, ce qui permet aux processus de datamining de fonctionner correctement.
L'existence de très nombreuses recherches autour du sujet prouve, si besoin était, l'importance du phénomène et ses améliorations constantes.

N'oublions pas que ce domaine est extrêmement récent : pourtant, il capitalise déjà un nombre immense de publications et des avancées se font à un rythme presque quotidien.
Les outils mathématiques s'affûtent, gagnant en complexité afin de permettre de traiter des volumes de données toujours plus grands. Les théories de prédiction et d'analyse deviennent de plus en plus fiables et se démocratisent, permettant à tous de profiter de la richesse des informations contenues.

Mais si les scientifiques avancent à grands pas, les experts marketing ne sont pas en reste. Les théories et les avancées sont nombreuses dans ce domaine aussi : voilà quelques années, le concept même de marketing viral n'existait pas, alors qu'il est aujourd'hui devenu l'un des fers de lance de la communication par Internet.

Malheureusement, les problématiques de vie privées évoluent au moins aussi vite. La somme d'informations qu'un individu est prêt à mettre en ligne est presque sans limite, contrairement à sa volonté de les partager avec le monde. Échanger avec un cercle contrôlé, oui ; avec le monde entier ou le gouvernement, non. Ces attentes doivent être prises en compte par les différents réseaux lors de l'exploitation des données grâce aux différents mécanismes désormais mis en place.

Comme souvent, au vu de l'état de l'art dans le domaine, une grande question demeure : qu'en sera-t-il dans dix ans ? Quelle problématique l'emportera, les mathématiques, le marketing ou le repli face au public ? La question en elle-même manque de pertinence, tant les concepts présentés ici sont neufs et risquent d'évoluer en quelques mois seulement.

A Modélisation des réseaux sociaux

> « *MySpace is doubly awkward because it makes public what should be private. It doesn't just create social networks, it anatomizes them. It spreads them out like a digestive tract on the autopsy table. You can see what's connected to what, who's connected to whom.* » *[TI06]*

L'analyse des réseaux sociaux est une problématique assez ancienne qui n'est pas née avec l'avènement des réseaux sociaux modernes tels que Facebook ou Twitter. En effet, l'analyse des données issues des relations sociales a occupé depuis plusieurs décennies de nombreux domaines tels que la sociologie, la psychologie ou encore l'anthropologie. Ces relations autrefois transparentes sont désormais rendues visibles grâce à l'informatique.

On pourra notamment citer les travaux de [Zac77], suivant les relations existantes entre les 34 membres d'un club de karaté sur une période de 3 ans.

On trouve notamment dans cet ouvrage un exemple de modélisation en graphe des réseaux sociaux, les *individus* étant représentés par des *nœuds*, les *liens* entre ces individus étant représentés par des arêtes, lesdits nœuds et lesdites *arêtes* pouvant avoir des attributs spécifiques.

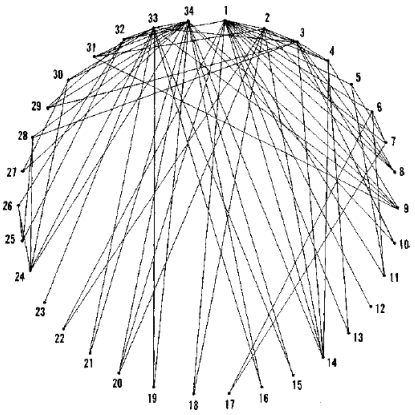

FIGURE 11 – [Zac77] Représentation graphique des relations sociales entre les 34 membres du club de karaté. Un lien existe entre deux membres si ceux-ci entretiennent des relations sociales en dehors du contexte du club de karaté.

Au cours de cette étude s'est déroulé un phénomène inattendu qui a été la division de ce club en deux à cause de divergences sur l'organisation du

club. Zachary a alors remarqué que cette division s'est caractérisée par une coupure du graphe représentatif des membres du club par sa coupe minimale, séparant ainsi les personnes ayant des opinions différentes sur le problème [11]. Cette particularité illustre l'existence d'une clusterisation de ces graphes : les personnes se groupent en *communautés*. Cette expérimentation a aussi mis en évidence la notion de *hubs*, sur laquelle nous reviendrons plus tard.

L'exemple donné ne traite que d'un graphe à petite échelle mais nous verrons qu'historiquement, deux axes de recherches sont apparus, pour l'étude de petits graphes par les scientifiques du domaine du social et par l'étude des propriétés des grands graphes.

Aujourd'hui, il est beaucoup plus simple de récolter les données nécessaires pour réaliser les recherches. Celles-ci proviennent en général de trois catégories de sources : les réseaux sociaux (MySpace, Facebook, LiveJournal, Flickr, etc.), les réseaux de communication en ligne (e-mail et messagerie instantanée, blogs, e-commerce, etc.) et les réseaux de publications en ligne (arXiv, etc.).

Contextualisons pour donner des exemples de dimensionnement. Dans les années 1970, l'étude portait sur des graphes de 34 nœuds sur 3 ans issus d'un club de karaté [Zac77]. En 2005, ce sont 436 nœuds issus d'échanges mails entre chercheurs des HP Labs qui sont étudiés sur 3 mois [AA05], puis 43 553 nœuds issus d'échanges mails sur une université pendant 1 an [KW06]. D'autres recherches ont également portés sur des grands réseaux comme les travaux de [LNNK+05] traitant un réseau de 1 312 454 nœuds ou des travaux de [LH07] qui ont révolutionné le domaine sur 240 millions d'individus, issus du réseau Microsoft Messenger.

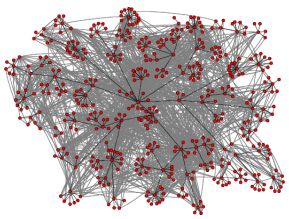

FIGURE 12 – [AA05] Réseau d'échange mails entre les membres des HP Labs, les liens gris étant les liens hiérarchiques et les liens gris les échanges mails.

11. Ce résultat est bien évidemment spécifique à ce cas là et ne peut être généralisé sans précautions.

Il est aussi à noter que la taille de ces graphes pose actuellement des problèmes en matière de stockage, surtout si ceux-ci sont dynamiques, jusqu'à représenter 100 GB de données par jour selon [Hof09] pour des graphes de 10^7 nœuds, 10^2 liens par nœuds et en prenant les métadonnées rattachées aux nœuds et aux liens ainsi que le dynamisme du graphe.

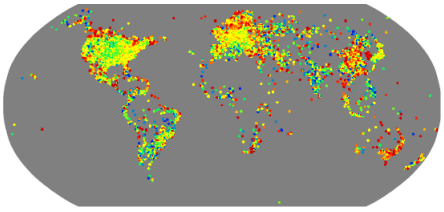

FIGURE 13 – [LH07] Utilisateurs du réseau Microsoft Messenger représentés géographiquement. La couleur est représentative de la densité à une position donnée.

Nous pouvons dès lors aller plus loin et essayer, dans un premier temps, de caractériser ces graphes pour ensuite essayer de retrouver ces comportements dans des modèles.

A.1 Propriétés des réseaux sociaux

A.1.1 Vocabulaire

Réseau social

Structure par laquelle des individus sont liés entre eux par un lien. Un tel réseau est généralement représenté par un graphe dont les nœuds sont les acteurs du réseau et dont les liens illustrent les relations entre ces acteurs.

Communautés, Clusters et Groupes

Une *communauté* est un ensemble de nœuds ayant de nombreuses connexions entre eux mais ayant un très faible nombre de connexions vers l'extérieur.

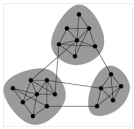

FIGURE 14 – Communautés dans un graphe.

Les communautés, aussi appelées *modules*, ne sont pas à confondre avec les *groupes*, ces derniers étant une caractéristique des nœuds. En effet, chaque nœud peut déclarer qu'il appartient à un ou plusieurs groupes (comme c'est le cas avec les groupes Facebook par exemple). Un groupe est donc constitué de membres adhérents et ne doit donc pas être confondu avec les communautés.

Enfin, la différence entre la détection de communautés et le clustering consiste en ce que le premier cherche à diviser le graphe en des structures selon leurs connexions (c'est uniquement de la topologie) alors que le deuxième consiste en un regroupement de structures selon une mesure de similarité, mesure qui est à définir au préalable. Cependant, ces deux notions sont similaires et sont souvent confondues.

A.1.2 Loi de distribution des degrés

Tout d'abord, il faut savoir que les études sur les graphes de réseaux sociaux ont montré que ces graphes possèdent des caractéristiques particulières. L'une de ces caractéristiques est la loi de distribution des degrés de ces graphes.

En effet, il a été observé par [Pri65] [12] puis vérifié plus tard avec expérimentations que dans un graphe représentant les réseaux sociaux, les degrés des nœuds suivent une distribution de type *loi de puissance*, de type $P(k) \propto k^{-\alpha}$ avec k les degrés d'un nœud.

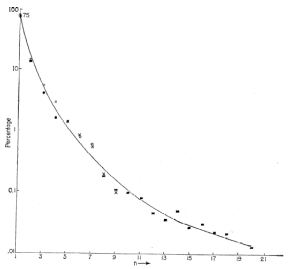

FIGURE 15 – [Pri65] Distribution des degrés obtenue avec en abscisses k et en ordonnées P(k).

Ces graphes sont alors dits à « invariance d'échelle » et sont bien différents de la répartition classique des degrés d'un graphe totalement aléatoire qui,

12. En l'occurrence, l'étude de [Pri65] est basée sur les citations entre les documents de recherche de l'époque.

dans un tel cas, suit une distribution binomiale de type $P(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

D'une façon plus générale, les lois de puissance se retrouvent souvent dans les réseaux d'informations comme les degrés des routeurs Internet. On parle d'une distribution *heavy-tailed* dont l'expression vérifie $\lim_{x\to\infty} e^{\lambda x} Pr[X > x] = \infty, \forall \lambda > 0$.

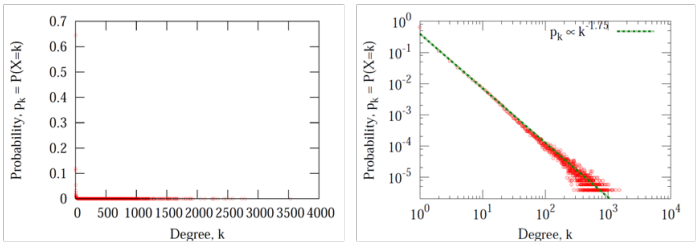

FIGURE 16 – [Les08] Distribution des degrés dans le réseau social Flickr. Le graphique de droite reprend les données de celui de gauche sur des axes en log-log. Il y a bien une loi de puissance avec ici un coefficient $\alpha = 1,75$.

L'une des conséquences de cette loi de distribution des degrés est qu'environ 20 % des nœuds ont 80 % des liens, donnant alors une importance capitale aux *hubs* (nœuds ayant beaucoup de liens sortants) et aux autorités (nœuds ayant beaucoup de liens entrants). En effet, la conséquence de ce type de propriété est qu'une attaque sur seulement 5 % des nœuds préalablement bien choisis suffit à défaire le réseau selon [Jac09].

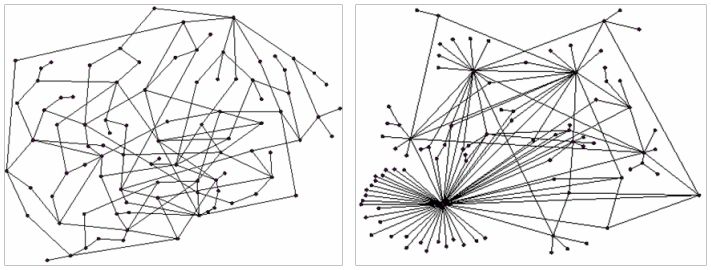

FIGURE 17 – Différence entre un graphe aléatoire (avec une loi de distribution des degrés binomiale) et un graphe représentant un réseau social. On notera la présence de nœuds spéciaux appelés « hubs » et « autorités » dont on reviendra dans la suite de ce document.

A.1.3 Diamètre et « petit monde »

En plus de la loi de distribution des degrés, il a été vérifié par [TM69] que les graphes ont, en général, un faible *diamètre*, qui rappelons-le est la distance la plus longue de toutes les plus courtes distances d'un graphe. Cette théorie est plus connue sous le nom des « 6 degrés de séparation » et caractérise un réseau dans lequel il est possible de trouver un chemin très court entre toute paire de points, chemin trouvable sans même connaître le réseau dans sa totalité. En général, le diamètre d'un graphe représentant un « petit monde » ou d'un graphe aléatoire est de l'ordre de $O(\log n)$ avec n le nombre de nœuds du graphe.

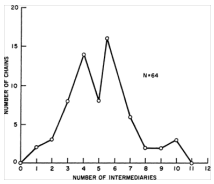

FIGURE 18 – [TM69] Longueur des chaînes complétées lors de l'expérience de Milgram, la moyenne étant ici de 6,2 étapes avant la réception finale.

Cette théorie se retrouve plus ou moins avec de vrais graphes tels que celui présenté par [LH07], présentant la répartition des distances sur le réseau Microsoft Messenger et qui représentée ci-dessous.

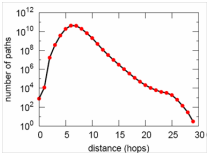

FIGURE 19 – [LH07] « Degrés de séparation » observés sur le réseau Microsoft Messenger. La moyenne est ici de 6,6 étapes.

Une autre étude de [MMG+07] nous montre qu'en 2006/2007, le réseau YouTube avait un diamètre de 21 (1,1 millions d'utilisateurs, 4,9 millions de liens, longueur moyenne des chaînes de 5,10), que le réseau Flickr avait un diamètre de 27 (1,8 millions d'utilisateurs, 22 millions de liens, longueur moyenne des chaînes de 5,67) et que le réseau LiveJournal avait un diamètre de 20 (5,2 millions d'utilisateurs et 72 millions de liens, longueur moyenne des chaînes de 5,88).

A.1.4 Dynamique et évolution

Tout d'abord, en opposition à ce qui se croyait auparavant, [LKF05] ont observé qu'a l'opposé de graphes « normaux », les graphes des réseaux sociaux devenaient de plus en plus denses avec l'augmentation des degrés des nœuds des graphes selon la loi de puissance suivante : $E(t) \propto N(t)^a$ avec $N(t)$ les nœuds au temps t, $E(t)$ les liens au temps t et a l'exposant de densification, $1 \leq a \leq 2$.

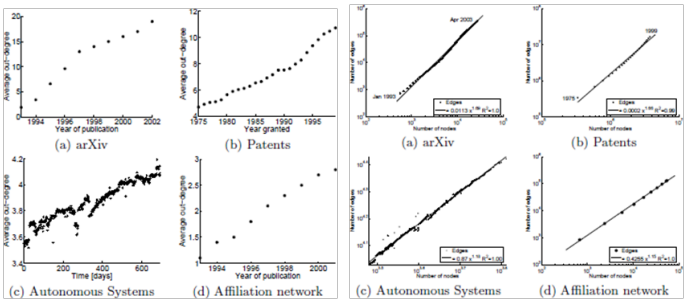

FIGURE 20 – [LKF05] À gauche, augmentation du degré des nœuds et à droite, corrélation avec la loi de densification proposée. Cette étude a été réalisée à partir de quatre sources de données.

On pourra remarquer que la loi de densification présentée suppose que le nombre de liens augmente plus vite que le nombre de nœuds. Enfin, ceux-ci ont également observés une *diminution du diamètre du graphe*.

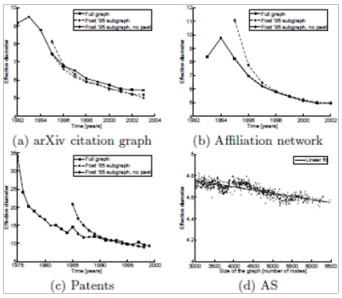

FIGURE 21 – [LKF05] Observation de la diminution du diamètre des graphes étudiés.

A.1.5 Clusterisation

Un réseau social se démarque également par son haut taux de clusterisation, qui peut se mesurer par la proportion de « mes » amis qui sont amis entre eux. Ce taux, pour un nœud i est calculé de la manière suivante : $C_i = \frac{2e_i}{k_i(k_i-1)}$, $C_i \in [0,1]$ avec e_i le nombre de liens entre les voisins de i et avec k_i le degré de i. Le taux de clusterisation peut également être calculé pour un graphe en effectuant la moyenne de tous les coefficients de chaque nœud, $C = \frac{1}{N}\sum_i^N C_i$.

Ce haut taux se retrouve dans les études de [New03] et de [New01] à partir desquelles nous savons que pour un réseau d'auteurs d'articles de recherche en physique constitué de 52 909 nœuds le taux de clusterisation est de 56 %. Il en est de même pour un réseau d'acteurs de cinéma pour lequel ce taux atteint les 78 %. Ce clustering peut s'apparenter à des communautés d'individus.

	network	type	n	m	z	ℓ	α	$C^{(1)}$	$C^{(2)}$	r	Ref(s).
social	film actors	undirected	449 913	25 516 482	113.43	3.48	2.3	0.20	0.78	0.208	20, 416
	company directors	undirected	7 673	55 392	14.44	4.60	–	0.59	0.88	0.276	105, 323
	math coauthorship	undirected	253 339	496 489	3.92	7.57	–	0.15	0.34	0.120	107, 182
	physics coauthorship	undirected	52 909	245 300	9.27	6.19	–	0.45	0.56	0.363	311, 313
	biology coauthorship	undirected	1 520 251	11 803 064	15.53	4.92	–	0.088	0.60	0.127	311, 313
	telephone call graph	undirected	47 000 000	80 000 000	3.16		2.1				8, 9
	email messages	directed	59 912	86 300	1.44	4.95	1.5/2.0		0.16		136
	email address books	directed	16 881	57 029	3.38	5.22	–	0.17	0.13	0.092	321
	student relationships	undirected	573	477	1.66	16.01	–	0.005	0.001	−0.029	45
	sexual contacts	undirected	2 810				3.2				265, 266
information	WWW `nd.edu`	directed	269 504	1 497 135	5.55	11.27	2.1/2.4	0.11	0.29	−0.067	14, 34
	WWW Altavista	directed	203 549 046	2 130 000 000	10.46	16.18	2.1/2.7				74
	citation network	directed	783 339	6 716 198	8.57		3.0/–				351
	Roget's Thesaurus	directed	1 022	5 103	4.99	4.87	–	0.13	0.15	0.157	244
	word co-occurrence	undirected	460 902	17 000 000	70.13		2.7		0.44		119, 157
technological	Internet	undirected	10 697	31 992	5.98	3.31	2.5	0.035	0.39	−0.189	86, 148
	power grid	undirected	4 941	6 594	2.67	18.99	–	0.10	0.080	−0.003	416
	train routes	undirected	587	19 603	66.79	2.16	–		0.69	−0.033	366
	software packages	directed	1 439	1 723	1.20	2.42	1.6/1.4	0.070	0.082	−0.016	318
	software classes	directed	1 377	2 213	1.61	1.51	–	0.033	0.012	−0.119	395
	electronic circuits	undirected	24 097	53 248	4.34	11.05	3.0	0.010	0.030	−0.154	155
	peer-to-peer network	undirected	880	1 296	1.47	4.28	2.1	0.012	0.011	−0.366	6, 354
biological	metabolic network	undirected	765	3 686	9.64	2.56	2.2	0.090	0.67	−0.240	214
	protein interactions	undirected	2 115	2 240	2.12	6.80	2.4	0.072	0.071	−0.156	212
	marine food web	directed	135	598	4.43	2.05	–	0.16	0.23	−0.263	204
	freshwater food web	directed	92	997	10.84	1.90	–	0.20	0.087	−0.326	272
	neural network	directed	307	2 359	7.68	3.97	–	0.18	0.28	−0.226	416, 421

TABLE II Basic statistics for a number of published networks. The properties measured are: type of graph, directed or undirected; total number of vertices n; total number of edges m; mean degree z; mean vertex–vertex distance ℓ; exponent α of degree distribution if the distribution follows a power law (or "–" if not; in/out-degree exponents are given for directed graphs); clustering coefficient $C^{(1)}$ from Eq. (3); clustering coefficient $C^{(2)}$ from Eq. (6); and degree correlation coefficient r, Sec. III.F. The last column gives the citation(s) for the network in the bibliography. Blank entries indicate unavailable data.

FIGURE 22 – [New03] Quelques données sur divers réseaux sociaux.

Cette propriété de clustering peut être expliquée par la propriété de similarité (les amis de x lui sont similaires) et par une transitivité de la similarité

où si x et y sont amis avec u alors x et u sont similaires, de même entre y et u d'où par transitivité x et y sont similaires ou encore x et y sont amis. Cette propriété est appelée *homophily*.

A.2 Modélisation des réseaux sociaux

> « *We want Kepler's Laws of Motion for the Web* »*(Mike Steuerwalt, NSF KDI Workshop, 1998)*

Les résultats détaillés précédemment sur la modélisation des réseaux sociaux ont été le fruit des recherches des dix dernières années. À l'heure des réseaux sociaux numériques, de nouveaux problèmes se posent, ce que nous allons voir dans la partie qui suit. Deux grands axes de recherche sont distinguables :

- Étude de la structure des réseaux sociaux (ce que nous allons traiter dans cette partie) ;
- Étude des propriétés et des motifs retrouvables dans les réseaux sociaux (traités dans la partie suivante).

Les techniques habituellement utilisées pour faire de la Fouille de Données ne sont pas adaptées à la nature même des graphes représentant les réseaux sociaux. En effet, de nombreux obstacles rendent ces algorithmes inefficaces dont notamment ceux de la taille des graphes soumis (qui ne cesse d'augmenter de par la facilité d'acquisition des données) et de la dynamique de ces graphes.

De plus, comme nous venons de le mentionner, l'aspect dynamique de ces réseaux sociaux soulève des problématiques jusqu'alors non abordées.

Les modèles présentés dans la suite de cette synthèse ont tous été conçus dans le but de reproduire les mécanismes régissant les propriétés observées précédemment et ce afin de pouvoir « prédire » d'autres propriétés jusqu'alors insoupçonnées.

A.2.1 Modélisation de la structure d'un réseau social par génération aléatoire

Des modèles de réseaux sociaux ont été élaborés dans le but d'améliorer la compréhension globale de ces réseaux. Parmi ceux-ci nous pourrons citer celui de [ER60] qui est généré aléatoirement [13]. Un modèle aussi simple qui

13. Les liens du graphe sont créés entre des nœuds choisis aléatoirement.

celui-là permet effectivement d'obtenir l'existence de courts chemins mais ne permet pas d'obtenir des degrés suivant la bonne distribution de degrés ni de retrouver une clusterisation caractéristique des petits mondes.

En effet, nous pouvons rappeler que les graphes obtenus de cette manière ont des degrés suivant une distribution de type binomiale ce qui ne correspond pas à ce qui est caractéristique aux réseaux sociaux, même si leur diamètre est bien de l'ordre de $O(\log n)$.

A.2.2 Modélisation de la structure d'un réseau social par les « petits mondes »

Il a fallu attendre les modèles de [WS98] générant des graphes avec de la clusterisation et un petit diamètre ainsi que le modèle de [Kle00] qui permettent tous les deux de retrouver algorithmiquement les caractéristiques des petits mondes sur des graphes générés, c'est-à-dire les propriétés de petit diamètre et un taux de clusterisation élevée.

Modèle de Watts & Strogatz

Le modèle de [WS98] est présenté ci-dessous. Celui-ci consiste simplement, à partir d'un graphe régulier (un nœud est relié à ses k voisins), de parcourir chacun des nœuds de ce graphe et de choisir, pour chaque nœud, un lien que l'on va, ou non, reconnecter avec une autre extrémité selon une probabilité p à définir.

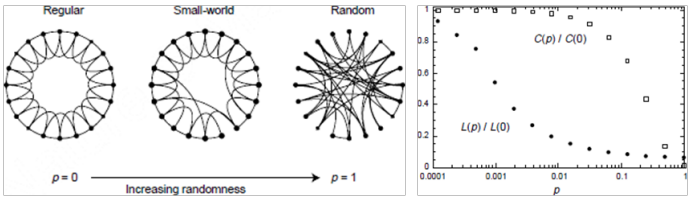

FIGURE 23 – [WS98] À gauche, la construction du graphe selon le choix de p. À droite, l'équilibre entre la longueur des cours chemins L du graphe et son taux de clusterisation C. Pour rappel, p doit être choisi tel que L soit petit et que C soit le plus grand possible.

En effet, comme nous pouvons le voir sur la figure ci-dessus, les réseaux sociaux sont considérés comme ayant une structure à partir de laquelle on ajoute une perturbation aléatoire paramétré avec un certain coefficient p.

De tels graphes ne sont cependant pas considérés comme des « petits mondes » puisque leur propriété de navigation n'est que très difficilement réalisable.

Modèle de Kleinberg

L'idée de [Kle00] est de disposer n individus sur une grille de dimension k et de connecter chaque individu avec, d'une part, leurs voisins les plus proches mais aussi avec d'autres personnes éloignées avec la probabilité de $Pr[u \to v] \propto \frac{1}{d(u,v)^\alpha}$ si on observe le lien entre un nœud u et un nœud distant v.

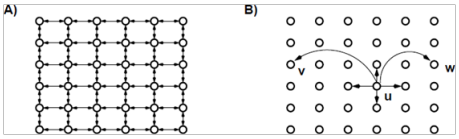

FIGURE 24 – [Kle00] Construction d'un graphe selon le modèle présenté.

Kleinberg a démontré que si $\alpha \neq k$, alors il n'existe pas d'algorithme local permettant de trouver des courts chemins mais que si $\alpha = k$ (la dimension de la grille) alors il est possible de trouver de tels chemins en utilisant l'algorithme glouton suivant : si on veut atteindre une cible c alors on doit choisir d'aller vers notre ami qui aime le plus c. Ces chemins ont alors une longueur de $O(\log^2 n)$. Cette différence d'avec le modèle de [KW06] permet de retrouver l'effet des petits mondes : ces graphes sont dits *routables*.

Ces algorithmes de routage sont actuellement exploités dans les réseaux d'échanges Peer-to-Peer, comme celui de [CSWH00] qui est utilisé dans le réseau Freenet.

Adaptations de l'approche de Kleinberg

Même si l'approche de Kleinberg permet effectivement d'obtenir un effet de petit monde, celle-ci a échouée dans la modélisation des données d'un vrai réseau social tel que celui de LiveJournal comme l'ont montrés [LNNK+05]. Ces derniers ont effectivement montrés que la répartition en $Pr[u \to v] \propto d(u,v)^{-2}$ (la surface de la Terre est de dimension 2) n'est pas vérifiée, le théorème de Kleinberg ne se basant que sur une grille de répartition des nœuds uniformes.

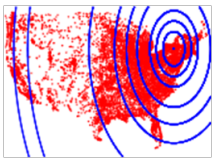

FIGURE 25 – [LNNK+05] Non uniformité des positions géographiques des utilisateurs du réseau social LiveJournal. Les cercles bleues représentent des paliers de 50 000 individus.

L'aspect géographique des individus représentés par les nœuds du réseau a donc son importance, importance retrouvée dans les approches faisant intervenir les rangs entre deux nœuds en lieu et place des distances. On a alors $Pr[u \to v] \propto \frac{1}{rank_u(v)}$ avec $rank_u(v) = |\{n : d(u,n) < d(u,v)\}|$ et qui représente donc le nombre de nœuds séparant u et v. Il a été démontré que cette approche basée sur le rang et donc sur la *densité* permet d'obtenir, de la même manière que celle de [Kle00], des petits mondes avec des chemins de longueur en $O(\log^3 n)$ cette fois-ci.

Il est également à noter que cette recherche a abouti sur une estimation de la probabilité que u et v soient amis suivante : $Pr[u \to v] = \epsilon + f(d(u,v))$ où $f(d(u,v))$ représente les amis géographiquement proches et où ϵ représentent les autres amis éloignés. Dans le cas de LiveJournal, cet ϵ représente 33 % des amis ce qui est non négligeable.

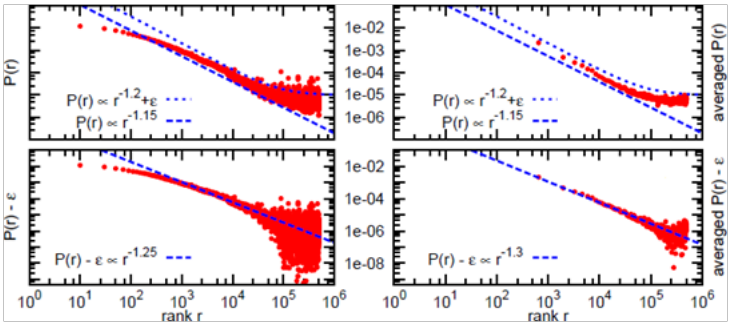

FIGURE 26 – [LNNK+05] Relation entre la probabilité d'être amis et le rang. On remarquera que la loi donnée ci-dessus avec les rangs est bien respectée.

Autres approches « petits mondes »

D'autres approches similaires à celles abordées précédemment ont été étudiées, toutes utilisant un graphe augmenté noté (H, ϕ), avec H un graphe et $\phi_u(v)$ la probabilité qu'un nœud v soit en lien avec un nœud u.

Autres approches issues d'autres domaines scientifiques

Pour sortir de l'approche « petits mondes », de nombreux autres modèles ont également été conçus pour représenter les réseaux sociaux. Ces modèles sont historiquement issus d'autres disciplines comme des Statistiques (p* Models [WP96], Stochastic Actor Oriented Model [SSS05], etc.), de la Sociologie, de la Physique (modèle d'Albert et de Barabási [AB02], etc.) ou des Mathématiques.

A.2.3 Modélisation de la dynamique et de l'évolution

Maintenant que des modèles sont capables de reproduire la structure globale d'un graphe représentant un réseau social, nous pouvons nous intéresser au côté dynamique derrière ces réseaux et notamment à ce qui se passe lors de l'ajout d'un nouveau nœud dans le graphe.

Approche guidée par la préférence

Dans le but d'obtenir une distribution des degrés qui suit une loi de puissance, les nouveaux nœuds arrivants ne peuvent pas avoir des degrés choisis sans précautions.

Un modèle intéressant de [Pri65] propose de réaliser des graphes tels que la probabilité d'avoir un lien entre un nouveau nœud u et un nœud existant v est lié au degré de la manière suivante : $Pr[u \to v] = \frac{deg(v)}{\sum_w deg(w)}$ avec w les autres nœuds du graphe, v y compris.

Ce modèle est souvent appelé « le phénomène du riche qui devient de plus en plus riche » et permet bien de garder une loi de distribution des degrés en loi de puissance. Cette analogie peut être étendue aux travaux de recherches puisqu'effectivement, comme l'observe [Pri65], les chances de nouvelles citations d'un papier de recherches sont proportionnelles aux citations qu'il possède déjà.

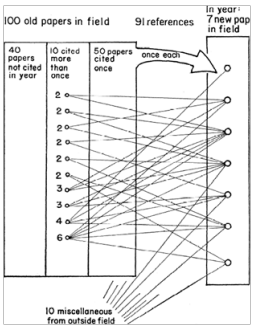

FIGURE 27 – [Pri65] Observations quant à la chance d'avoir de nouvelles citations selon les citations déjà possédées.

D'autres modèles, tous basés sur le principe du guidage par la préférence, ont été conçus pour compléter ce modèle qui ne prend pas en compte plusieurs paramètres comme la topologie du graphe par exemple.

Modèle du feu de forêt

Le modèle dit du feu de forêt a été conçu par [LH07]. Celui-ci est très simple mais permet de modéliser la dynamique d'un graphe de réseau social en le densifiant tout en réduisant son diamètre.

L'algorithme est le suivant : pour chaque nouveau nœud v qui arrive, choisir au hasard (loi uniforme) un nœud w qui servira « d'ambassadeur ». À partir de ce nœud, déterminer au hasard un nombre de connexions entrantes et sortantes de w et répéter ces étapes jusqu'à ce que le « feu » s'arrête. Cette approche est basée sur le principe que chaque nouveau nœud dispose d'un centre de gravité et que la probabilité que celui-ci soit liés à d'autres nœuds décroît selon la distance à laquelle les autres nœuds ce trouvent par rapport à ce centre. Le nouveau nœud v est alors attaché à tous les nœuds ayant été touchés par le feu.

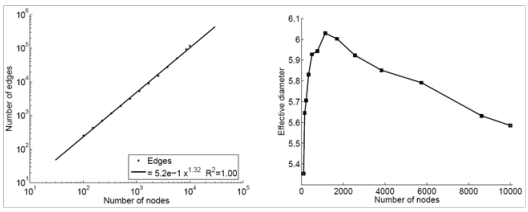

FIGURE 28 – [LKF05] Densification et réduction du diamètre d'un graphe après l'ajout de nouveaux nœuds selon l'algorithme du feu de forêt.

Modèle basé sur les produits de Kronecker

Un nouveau modèle assez récent proposé par [LCKF05] et par [LCK+10] permet de générer des graphes ayant des sous-graphes se répétant. En effet, les observations font part de parties de graphes qui se répètent d'où l'utilisation du produit de Kronecker pour générer de tels graphes. Pour rappel, le produit de Kronecker (noté $\otimes$) s'exprime de la façon suivante :

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}$$

Le résultat d'un produit de Kronecker peut alors s'assimiler à des matrices d'adjacences, dont la taille est continuellement en augmentation. Un graphe de Kronecker est alors défini de la manière suivante : $K_1^{[k]} = K_k = K_1 \otimes K_1 \otimes ... \otimes K_1 = K_{k-1} \otimes K_1$. La matrice K_1 sert à initialiser la génération et est dite matrice d'initialisation.

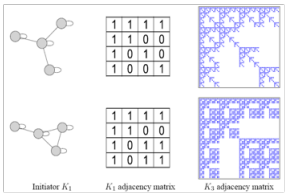

FIGURE 29 – [LCK+10] Exemple de génération de matrices de Kronecker avec des matrices d'initialisations différentes.

Le modèle de Kronecker nous permet de générer des graphes vérifiant la plupart des caractéristiques listées précédemment.

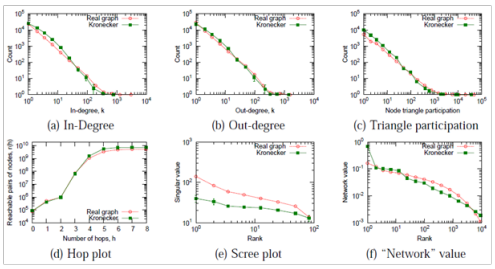

FIGURE 30 – [LCK+10] Corrélation entre l'évolution du modèle basé sur le produit de Kronecker avec ce qui est observé.

Le modèle de Kronecker nous offre un cadre permettant de nombreuses possibilités puisque celui-ci permet de faciliter les simulations d'algorithmes, de faire des prédictions, de détecter des anomalies et bien d'autres applications.

Références

[AA01] D. Agrawal and C. C. Aggarwal, 2001. URL : http ://charuaggarwal.net/private.pdf.

[AA05] Adamic and Adar. How to search a social network, 2005.

[AB02] Albert and Barabási. Statistical mechanics of complex networks, 2002.

[AES03] Johannes Gehrke Alexandre Evfimievski and Ramakrishnan Srikant. *ACM*, 2003. URL : http ://www.cs.cornell.edu/johannes/papers/2003/pods03-privacy.pdf.

[Ast04] Philippe Astor. Un groupe de défense de la vie privée porte plainte contre le service gmail de google. *ztndet.fr*, 2004. URL : http ://www.zdnet.fr/actualites/un-groupe-de-defense-de-la-vie-privee-porte-plainte-contre-le-service-gmail-de-google-39148165.htm.

[Bar06] Michael Barbaro. A face is exposed for aol searcher no. 4417749. *The New York Times*, August 2006. URL : http ://www.nytimes.com/2006/08/09/technology/09aol.html ?pagewanted=all.

[Bou05] Marie Boulain. Marketing de masse ou marketing direct. 2005. URL : http ://www.fcsq.qc.ca/publications/savoir/Decembre-2005/Page-6.pdf.

[BP99] Brin and Page. The anatomy of a large-scale hypertextual web search engine, 1999.

[CMN08] Clauset, Moore, and Newman. Hierarchical structure and the prediction of missing links in networks, 2008.

[CNM04] Clauset, Newman, and Moore. Finding community structure in very large networks, 2004.

[CSWH00] Clarke, Sandberg, Wiley, and Hong. Freenet : A distributed anonymous information storage and retrieval system, 2000.

[Dem97] A.P. Dempster. *Journal of the Royal Statistical Society*, 1997.

[DKT03] Jon Kleinberg David Kempe and Eva Tardos. Maximizing the spread of in[U+FB02]uence through a social network, 2003. URL : http ://www.cs.cornell.edu/home/kleinber/kdd03-inf.pdf.

[DN04] Irit Dinur and Kobbi Nissi. URL : http ://www.cs.bgu.ac.il/ kobbi/papers/psd.pdf, 2004.

[Dom04] Pedro Domingos. Mining social networks for viral marketing, 2004. URL : http ://dl.dropbox.com/u/18765753/mrk6092/9.viral-marketing/10.1.1.76.4474.pdf.

[DR01] Pedro Domingos and Matthew Richardson. Mining the network value of customers, 2001. URL : http ://www.cs.washington.edu/homes/pedrod/papers/kdd01a.pdf.

[DR02] Pedro Domingos and Matthew Richardson. Mining knowledge-sharing sites for viral marketing, 2002. URL : http ://www.cs.washington.edu/homes/pedrod/papers/kdd02b.pdf.

[Dum04] Estelle Dumout. Nouveau créneau pour google : la pub ciblée dans le courrier privé. *ztndet.fr*, 2004. URL : http ://www.zdnet.fr/actualites/nouveau-creneau-pour-google-la-pub-ciblee-dans-le-courrier-prive-39147650.htm.

[ER60] Erdõs and Rényi. On the evolution of random graphs, 1960.

[Fac11] Facebook. Brochure, 2011. URL : https ://www.facebook.com/press/info.php ?factsheet.

[GN01] Girvan and Newman. Community structure in social and biological networks, 2001.

[Goo11] Google. Google's income statement information, 2011. URL : http ://investor.google.com/financial/tables.html.

[Hof09] Hofman. Social network analysis with hadoop, 2009.

[Jac09] Jacomy. Introduction à l'exploration du web par la théorie des graphes, 2009.

[JLA07] Bernardo A. Huberman Jure Leskovec and Lada A. Adamic. The dynamics of voral marketing, 2007. URL : http ://www-personal.umich.edu/ ladamic/papers/viral/viralTWeb.pdf.

[Juv00] Steve Juvertson. What exactly is viral marketing ? 2000. URL : http ://currypuffandtea.files.wordpress.com/2008/03/viral-marketing.pdf.

[Kle97] Kleinberg. Authoritative sources in a hyperlinked environment, 1997.

[Kle00] Kleinberg. The small-world phenomenon : An algorithmic perspective, 2000.

[KW06] Kossinets and Watts. Empirical analysis of an evolving social network, 2006.

[Lau09] Samuel Laurent. Un internaute mis à nu à partir de ses traces sur le web. *Le Figaro*, January 2009. URL : http ://www.lefigaro.fr/hightech/2009/01/15/01007-20090115ARTFIG00625-un-internaute-mis-a-nu-a-partir-de-ses-traces-sur-le-web-.php.

[LCK+10] Leskovec, Chakrabarti, Kleinberg, Faloutsos, and Ghahramani. Kronecker graphs : An approach to modeling networks, 2010.

[LCKF05] Leskovec, Chakrabarti, Kleinberg, and Faloutsos. Realistic and mathematically tractable graph generation and evolution and using kronecker multiplication, 2005.

[Les08] Leskovec. Dynamics of large networks, 2008.

[LH07] Leskovec and Horvitz. Planetary-scale views on an instant-messaging network, 2007.

[LKF05] Leskovec, Kleinberg, and Faloutsos. Graphs over time : Densification laws and shrinking, 2005.

[LL98] Charles X. Ling and Chenghui Li. Data mining for direct marketing : Problems and solutions, 1998.

[LNK04] Liben-Nowell and Kleinberg. The link prediction problem for social networks, 2004.

[LNNK+05] Liben-Nowell, Novakz, Kumarz, Raghavanx, and Tomkinsz. Geographic routing in social networks, 2005.

[LTZ+07] Li, Tang, Zhang, Luo, Liu, and Hong. Eos : Expertise oriented search using social networks, 2007.

[LV07] Luxburg and V. A tutorial on spectral clustering, 2007.

[Met09] Raphaël Metz. Marc l***. *Le Tigre*, January 2009. URL : http ://www.le-tigre.net/Marc-L.html.

[MMG+07] Mislove, Marcon, Gummadi, Druschel, and Bhattacharjee. Measurement and analysis of online social networks, 2007.

[New01] Newman. Scientific collaboration networks, 2001.

[New03] Newman. Fast algorithm for detecting community structure in networks, 2003.

[NS06] Arvind Narayanan and Vitaly Shmatikov. Ancien titre : How To Break Anonymity of the Netflix Prize Dataset, URL : http ://arxiv.org/PS_cache/cs/pdf/0610/0610105v2.pdf, 2006.

[PDFV05] Palla, Derényi, Farkas, and Vicsek. Uncovering the overlapping community structure of complex networks in nature and society, 2005.

[PK09] Delphine Manceau Bernard Dubois Philip Kotler, Kevin Keller. *Marketing Management*. Pearson, 2009.

[Pri65] Price. Networks of scientific papers, 1965.

[PS09] Katherin Peterson and Katie A. Siek. Analysis of information disclosure on a social networking site. 2009.

[Pui11] Florian Puisais. La plateforme google+ contre facebook, ou l'enjeu des données personnelles pour la publicité online. *presse-citron.net*, 2011. URL : http ://www.presse-citron.net/la-plateforme-google-contre-facebook-ou-l%E2%80%99enjeu-des-donnees-personnelles-pour-la-publicite-online.

[SSS05] Snijders, Steglich, and Schweinberger. Modeling the co-evolution of networks and behavior, 2005.

[STLS05] Song, Tseng, Lin, and Sun. Expertisenet : Relational and evolutionary expert modeling, 2005.

[TI06] Tossell and I. Love to hate myspace ? check out the buzz, 2006.

[TM69] Travers and Milgram. An experimental study of the small world problem, 1969.

[WP96] Wasserman and Pattison. Logit models and logistic regressions for social networks : I. an introduction to markov graphs and p*, 1996.

[WS98] Watts and Strogatz. Collective dynamics of 'small-world' networks, 1998.

[Zac77] Zachary. An information flow model for conflict and fission in small groups, 1977.