L'analyse des réseaux sociaux

Françoise Fogelman Soulié [1] et Emmanuel Viennet [2]

1. Introduction

Depuis 2002, date de la création de MySpace, qui devint le premier site social mondial, les sites sociaux ont envahi notre vie : Facebook, créé en 2004, a atteint 1,44 milliards d'utilisateurs mensuels en mars 2015 [3], LinkedIn, créé en 2003, en a aujourd'hui environ 400 millions et Twitter (2006) compte 302 millions d'utilisateurs actifs. En France, Skyrock, créé en 2002, atteignait 21 millions d'utilisateurs en juin 2008 avant d'être atteint par la croissance de Facebook. En Chine, Tencent a 600 millions d'utilisateurs actifs par mois sur WeChat et 843 millions sur QQ, Sina Weibo 300 millions.

Parallèlement un nouveau domaine scientifique se développait : l'analyse des réseaux sociaux (ou SNA : *Social Network Analysis*). Le SNA trouve ses origines théoriques dans les travaux des mathématiciens sur les graphes [7], mais les premiers développements significatifs sont apparus en sciences sociales [19, 22] : les réseaux traités sont alors des réseaux d'individus reliés par des interactions sociales ou politiques et l'analyse, peu informatisée, ne peut en pratique traiter que des réseaux de relativement petite taille.

À partir des années 2000, en partie du fait du développement des moteurs de recherche et de l'e-commerce, les chercheurs, notamment en informatique, en *data*

1. School of Computer Software, Tianjin University, China.

2. Université Paris 13, Sorbonne Paris Cité, L2TI.

3. En croissance de 13 % sur 2014. Voir : *Facebook Reports First Quarter 2015 Results*. `http://investor.fb.com/releasedetail.cfm?ReleaseID=908022`

mining et en physique, commencent à s'intéresser aux systèmes organisés en réseaux. Un réseau est un ensemble d'entités reliées par des relations. Le titre du livre de Barabási [3], *Linked*: *How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*, illustre dès 2002 cette idée de réseau comme modèle conceptuel général de très nombreux phénomènes complexes. Le réseau peut alors représenter aussi bien un site social (comme Facebook), un réseau d'utilisateurs du téléphone ou du courrier électronique, le Web (un réseau de pages reliées par des liens hypertextes [12, 15]), un ensemble de consommateurs [16], de terroristes [20], d'auteurs d'articles scientifiques [22], de protéines en interaction [3], un réseau de distribution d'eau [17], etc. Le développement des sites sociaux a bien sûr encore renforcé l'activité du domaine de l'analyse des réseaux complexes [1, 2, 13].

Nous ne pouvons pas dans cet article présenter de façon exhaustive l'ensemble des travaux de ce domaine très foisonnant, qui pose de très intéressants problèmes informatiques, notamment liés aux méthodes de traitement de données de grandes dimensions connues sous l'appellation de *Big Data*. Nous allons essayer d'en illustrer quelques concepts, en nous appuyant, pour en faciliter la compréhension sans entrer dans les détails techniques (que le lecteur pourra approfondir dans les très nombreuses références fournies), sur les résultats de plusieurs projets collaboratifs auxquels nous avons participé : CADI (ANR : 2007-2010), CEDRES-ExDeuss (ANR-DGCIS : 2009-2013), eFraudBox (ANR : 2010-2013), AMMICO (FUI : 2012-2015), OFS (PSPC : 2012-2016), REQUEST (Appel à projets Cloud Computing No 3 « Big Data » : 2014-2018).

2. Qu'est-ce qu'un réseau social ?

Un *réseau social* représente un système d'entités en interaction. On le modélisera comme un graphe $G = (S, A)$ où S est un ensemble d'entités (les *sommets* ou nœuds du graphe) et A est l'ensemble des *arcs* (ou connexions) représentant les interactions entre ces sommets. Un site social comme Facebook peut également être représenté de cette façon : un nœud est un membre de Facebook, et deux nœuds sont connectés s'ils sont amis (voir Figure 1). L'analyse de ces données permet par exemple de construire un graphe montrant les interactions entre les utilisateurs des différents pays [4] : ici les liens sont d'autant plus foncés que le nombre d'échanges est important. La visualisation des réseaux sociaux est un domaine très actif sur lequel nous reviendrons plus loin.

Les entités d'un réseau social peuvent être de toutes sortes : pages web, membres d'un site social, comptes bancaires, protéines... et les liens peuvent représenter des interactions variées entre ces entités : liens d'amitié sur Facebook, *follower* sur Twitter, hyperliens sur le web, correspondance e-mail ou appels téléphoniques... On voit

4. http://www.emanueldeutschmann.net/home/category/social-network-analysis

FIGURE 1. Les relations internationales sur le réseau Facebook.

donc que de très nombreux exemples de « réseaux sociaux » n'ont rien de social et pourraient être appelés plus justement « réseaux complexes ». Mais le terme « réseau social » est dominant dans la littérature et c'est celui que nous utiliserons.

Un réseau peut être *dirigé* ou pas selon qu'on distingue l'interaction d'un sommet A vers B ou pas : par exemple une page Web A peut pointer vers une page B, alors que B ne pointe pas vers A.

Un réseau peut être *pondéré* ou pas selon qu'on attribue un poids aux liens, par exemple le nombre de messages envoyés du sommet A au sommet B dans la période de référence dans un réseau d'emails. Certains réseaux peuvent porter des étiquettes sur les liens plus complexes qu'un simple poids, pour décrire une superposition de réseaux partageant les mêmes sommets, mais avec des interactions de natures différentes (par exemple, le réseau des amis, des collègues de travail, des membres de la famille).

Outre leur très grande taille, les graphes représentant des réseaux sociaux sont très particuliers. Un réseau social possède une propriété de *localité* des relations : si A est relié à B et C, alors la probabilité que B et C soient eux-mêmes reliés est plus grande que si les relations étaient aléatoires : le graphe comporte beaucoup de triangles [22]. Cette propriété élémentaire n'est pas la seule qui caractérise les réseaux sociaux ; voici les plus importantes :

(1) Les réseaux sociaux sont souvent *très grands* : millions de nœuds, et millions, voire milliards de liens ;

(2) Il y a en général de nombreux *attributs* sur les nœuds (nom, adresse, possession d'un produit...), voire sur les liens (nombre d'appels téléphoniques entre les deux nœuds par exemple) ;

Françoise Fogelman-Soulié	coauthored with	Servet A. Martínez	MR1102355 (92f:92006)
Servet A. Martínez	coauthored with	Peter E. Ney	MR1804951 (2001k:37017)
Peter E. Ney	coauthored with	Paul Erdös	MR0373068 (51 #9270)

FIGURE 2. Nombre d'Erdös et propriété du petit monde.

(3) Les nœuds qui portent des attributs similaires ont tendance à être reliés : ce phénomène est nommé l'*homophilie* ;

(4) Le réseau peut souvent être décomposé en sous-groupes, ou *communautés*, où les nœuds sont très connectés entre eux, mais ont peu de liens avec les autres sous-groupes. Nous reviendrons plus loin sur cette notion ;

(5) Le réseau forme un *petit monde* : la longueur moyenne du plus court chemin entre deux nœuds est petite (la longueur du plus court chemin entre deux nœuds est la distance, ou *degré de séparation*, entre ces nœuds). Stanley Milgram [19] réalisa une expérience en 1967 pour démontrer cette règle du petit monde qu'on appelle aussi *règle des six degrés de séparation* : on remet une lettre à un échantillon de personnes dans quelques villes des États-Unis en leur indiquant le nom d'un destinataire à Boston. Si la personne connaît le destinataire personnellement, elle lui adresse directement la lettre. Sinon, elle cherche parmi ses amis une personne qui, à son avis, pourrait connaître personnellement le destinataire et elle lui envoie la lettre. Les résultats de l'expérience démontrèrent qu'il suffisait en moyenne de six envois pour atteindre le destinataire. Cette propriété a été exploitée dans le jeu des six degrés de Kevin Bacon (reliant un acteur à K. Bacon par le plus court chemin possible, deux acteurs étant reliés s'ils ont joué dans un même film) ou dans le nombre d'Erdös, reliant un scientifique à Erdös par le plus court chemin possible, deux chercheurs étant reliés s'ils ont été co-auteurs d'un même article scientifique. La figure 2 montre par exemple un nombre d'Erdös de 3 pour l'un des auteurs de cet article [5] : le réseau des collaborations scientifiques est un petit monde.

(6) Les graphes sociaux sont *sans-échelle* (*scale-free*) [3, 22]. On dit qu'un graphe est sans-échelle si la distribution des degrés suit une loi de Pareto (dite aussi loi de puissance), où le degré d'un nœud est le nombre de nœuds auxquels il est relié (γ est un nombre positif) :

$$P(\text{degré} = k) \approx k^{-\gamma}.$$

Ceci signifie que la plupart des nœuds sont peu connectés ($k = 1$ ou 2) mais qu'il est probable de rencontrer quelques nœuds très connectés, que l'on appellera des *hubs*. Ces nœuds sont essentiels dans la connectivité du réseau (penser au rôle de Roissy-Charles de Gaulle dans le réseau de transport aérien, au côté de dizaines de

5. http://www.ams.org/mathscinet/collaborationDistance.html

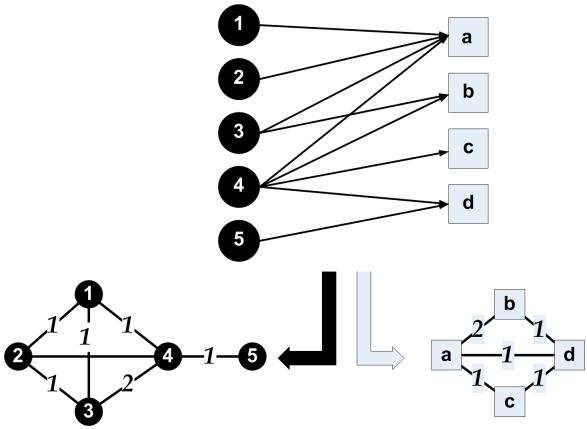

FIGURE 3. Un réseau bipartite (en haut) et les deux réseaux unipartites projetés (en bas), où on a indiqué le nombre k ou k' de voisins en commun.

petits aéroports peu connectés). Cette distribution très inégalitaire des degrés peut s'expliquer par des modèles de croissance du réseau comme l'*attachement préférentiel* [3].

Les réseaux dont nous venons de parler sont modélisés par des graphes « unipartites », formés d'une seule sorte de nœud. On utilisera dans la suite de cet article des *graphes bipartites* : un graphe bipartite est un graphe dans lequel les sommets sont répartis en deux familles disjointes, les liens ne pouvant joindre que des sommets de familles différentes. À partir d'un graphe bipartite, on peut construire deux graphes unipartites par « projection » : le premier réseau (resp. second réseau) a comme sommets les nœuds de la famille 1 (resp. de la famille 2). Deux sommets du graphe 1 (resp. du graphe 2) sont reliés s'ils sont connectés à au moins k (resp. au moins k') mêmes sommets du graphe 2 (resp. du graphe 1). Intuitivement, un graphe projeté relie des nœuds ayant des comportements similaires : par exemple, si le graphe bipartite représente des individus ayant acheté des produits, le graphe projeté des individus relie des individus ayant acheté au moins k produits identiques et celui des produits relie des produits ayant été achetés par au moins k' individus identiques. La projection est une étape qui peut être assez longue (complexité en $O(m \log n)$ où n est le nombre de nœuds et m le nombre d'arêtes du graphe bipartite), notamment si certains nœuds sont très connectés et ont donc beaucoup de voisins communs avec les autres nœuds (ces nœuds très connectés sont appelés « *mega-hubs* » et sont souvent responsables de la plupart des arêtes).

3. Les communautés

Qu'est-ce qu'une communauté ?

Un réseau social peut souvent être décomposé en communautés, groupes d'entités communiquant beaucoup entre elles. Plusieurs définitions formelles de cette notion ont été proposées, mais l'intuition reste la même : il s'agit de décomposer le réseau social en sous-groupes tels que :

(1) à l'intérieur d'un sous-groupe les nœuds soient très interconnectés, et

(2) il existe peu de liens entre deux sous-groupes.

Un sous-groupe devrait donc être proche d'une *clique* (c'est-à-dire un ensemble de nœuds totalement connectés).

La décomposition du réseau en communautés réalise donc une segmentation (*clustering* non supervisé en apprentissage statistique) des sommets. Alors que les techniques classiques de segmentation utilisent des mesures de *similarité* entre les nœuds, calculées à partir de leurs attributs, la détection de communautés est basée sur la structure du graphe sans prendre en compte les attributs (des méthodes hybrides considérant à la fois les attributs et les liens ont toutefois été proposées).

Les algorithmes de détection de communauté

Le partitionnement de graphes est un problème NP-difficile. Le nombre de partitions, B_n, d'un réseau de n nœuds croît plus vite qu'exponentiellement ; il y a par exemple 10^{40} partitions d'un réseau de $n = 50$ nœuds :

$$B_n = \frac{1}{e} \sum_{j=0}^{\infty} \frac{j^n}{j!}.$$

La détection de communautés est donc un problème difficile : les réseaux peuvent être très grands (le passage à l'échelle des algorithmes est donc un facteur critique), le nombre de communautés est inconnu, leurs tailles et densités peuvent être très différentes. Sans *vérité de terrain*, la validation des communautés détectées reste donc un problème largement mal résolu, comme c'est souvent le cas pour les méthodes d'apprentissage non supervisé, et en particulier le *clustering*.

Les algorithmes traditionnels de théorie des graphes, exacts mathématiquement, sont trop complexes pour traiter les très grands réseaux, pour lesquels seules des méthodes approchées en temps quasi-linéaire sont applicables. Les algorithmes de détection de communautés pour les réseaux se sont donc développés selon plusieurs axes, en particulier les suivants [4] :

— *Clustering hiérarchique.*
On définit une mesure de similarité entre les sommets ou les liens, basée sur la structure du réseau. Pour cela, on peut se baser sur des indices comme l'intermédiarité (le nombre de plus courts chemins passant par un nœud donné), l'index de Jaccard (défini plus loin) ou la distance de Hamming entre les lignes de la matrice d'adjacence du graphe. On regroupe alors les nœuds (algorithmes agglomératifs) ou on élimine les liens (algorithmes divisifs) ayant une forte similarité. Ces algorithmes ont une complexité en $O(n^2)$ (algorithme agglomératif de Ravasz), ou en $O(m^2 n)$ (algorithme divisif de Girvan et Newman, où m est le nombre de liens).

— *Méthodes d'optimisation de la modularité.*
Ces méthodes reposent sur l'hypothèse qu'un réseau aléatoire n'a pas de communauté.

La modularité a été introduite par Girvan et Newman en 2004 [11] : elle mesure l'écart entre le nombre de liens dans une communauté et ce que serait ce nombre dans un réseau aléatoire. En calculant le nombre de liens à l'intérieur de la communauté et vers l'extérieur de la communauté, on a :

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j).$$

où A_{ij} est le coefficient de la matrice d'adjacence du graphe en ligne i et colonne j ($A_{ij} = w_{ij}$ si les nœuds i et j sont connectés par un lien de poids w_{ij}, 0 sinon) ; k_i est la somme des poids des arêtes ayant i comme un de leurs sommets ; c_i est la communauté du nœud ; $\delta(u, v) = 1$ si $u = v$, 0 sinon et

$$m = \frac{1}{2} \sum_{i,j} A_{ij} = \frac{1}{2} \sum_{i} k_i = \frac{1}{2} \sum_{j} k_j$$

(dans le cas d'un graphe non pondéré, m est le nombre total d'arêtes).

Les algorithmes d'optimisation de la modularité sont basés sur des heuristiques gloutonnes, pour être praticables sur des grands réseaux [4]. L'un des plus efficaces aujourd'hui pour les grands graphes est l'algorithme dit *algorithme de Louvain* [5] qui peut traiter des réseaux de plusieurs centaines de millions de nœuds. C'est un algorithme itératif qui fonctionne en deux phases :

Phase 1 : après avoir initialisé les communautés (une par nœud), on considère, dans un ordre aléatoire, chaque nœud et on regarde si l'on peut augmenter la modularité Q en l'associant à la communauté de l'un de ses voisins. Si oui, le nœud est changé de communauté. On itère cette opération jusqu'à stabilisation.

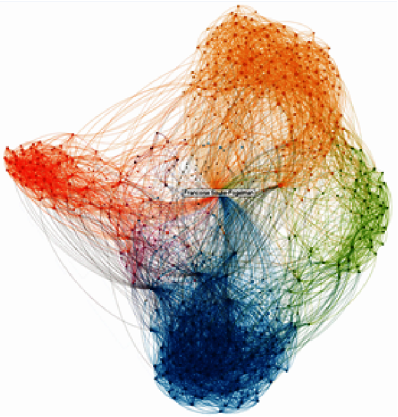

FIGURE 4. Réseau des liens LinkedIn et communautés extraites.

Phase 2 : on construit un nouveau graphe, dont les nœuds sont les communautés obtenues en phase 1. Un lien entre deux communautés a pour poids la somme des poids des liens de chaque nœud d'une communauté vers les nœuds de l'autre communauté.

On itère ces phases jusqu'à ce qu'on ne puisse plus augmenter significativement la modularité.

L'efficacité de l'algorithme vient de ce que le calcul du gain de modularité obtenu en déplaçant un nœud vers la communauté d'un de ses voisins se fait *localement* (phase 1). Par ailleurs (phase 2), l'algorithme produit une *décomposition hiérarchique* du réseau en communautés de plus en plus synthétiques, permettant d'explorer le réseau à des échelles différentes (voir Section 4). Cet algorithme est aujourd'hui utilisé très largement et c'est lui que nous utilisons dans toute la suite de cet article sauf mention contraire.

4. La visualisation

La visualisation des réseaux est un outil pour analyser les réseaux et identifier les caractéristiques critiques. Des logiciels *open source* de visualisation existent, mais sont souvent incapables de traiter les grands réseaux. Le logiciel Gephi [6] permet quant à lui de traiter des réseaux de plusieurs centaines de milliers de nœuds. La figure 4 montre par exemple le réseau LinkedIn d'un des co-auteurs de cet article (un « ego-réseau », c'est-à-dire centré sur l'individu et ne comportant que ses amis et leurs liens), représenté par l'application LinkedIn InMaps (arrêtée en 2014), où les

6. `https://gephi.github.io/`. Un autre logiciel libre a été développé par le LaBRI : Tulip (`http://tulip.labri.fr/`).

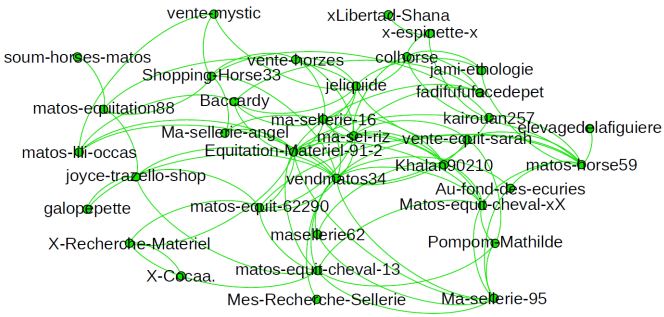

FIGURE 5. Communauté d'amis sur Skyrock intéressés par l'équitation.

communautés extraites par l'algorithme de Louvain (colorées chacune d'une couleur différente) représentent les différents groupes de collègues du parcours professionnel. La représentation graphique utilise Gephi.

Nous avons utilisé ce type de représentation pour analyser le réseau Skyrock [21], un réseau de blogs utilisé par les adolescents, `www.skyrock.com` (projet CEDRES-ExDeuss). Chaque membre choisit un pseudonyme et dispose d'un blog où il peut poster des messages, des articles, des images, des tags ; il peut se lier à des amis et recevoir des commentaires. L'ensemble de ces données représente pour Skyrock plus de 10 To chaque mois à manipuler [7]. Une des premières questions est d'identifier les « groupes d'intérêt » sur la plateforme. Pour cela on utilise les déclarations d'amis pour construire le *réseau des amis*, puis on calcule les communautés. Un traitement des pseudonymes, pour en extraire les n-grammes les plus fréquents, permet de fournir une caractérisation sémantique des communautés : la figure 5 montre par exemple une communauté de « skynautes » ayant choisi des pseudonymes en relation avec l'équitation.

Un travail du même genre, mais cette fois sur le réseau bipartite liant les individus aux tags, permet de construire des communautés d'intérêt, comme par exemple dans la figure 6 montrant deux communautés d'amateurs d'Adidas : l'une est composée d'amateurs de musique et de sport, et l'autre d'amateurs de vampires. Identifier de telles communautés permet aux opérateurs de campagnes de mieux cibler leurs bannières sur des groupes d'individus les plus susceptibles d'être intéressés par leurs bannières.

Dans un autre contexte, celui de l'analyse de la fraude à la carte bancaire sur Internet (projet eFraudBox), nous avons construit [8] un réseau bipartite liant les cartes aux marchands : une carte est liée à un marchand si elle y a effectué un achat. On

7. `http://www.lemondeinformatique.fr/actualites/lire-skyrock-dope-ses-forums-grace-a-un-outil-d-analyse-predictif-53036.html`

FIGURE 6. Tags associés à deux communautés d'intérêt autour de la marque Adidas.

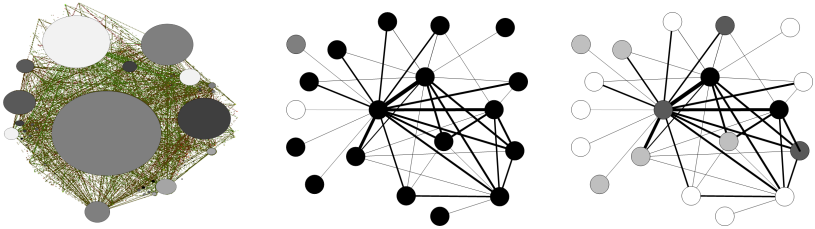

FIGURE 7. Communautés de marchands (à gauche). Puis une communauté de marchands colorée avec le montant de la fraude, au centre, et le taux de fraude, à droite.

peut ensuite calculer la décomposition hiérarchique en communautés du réseau des marchands, ce qui permet d'obtenir des visualisations comme dans la figure 7 avec Gephi. À gauche on représente le réseau des communautés de marchands : un nœud est une communauté, la taille du nœud est proportionnelle au nombre de marchands dans la communauté, la couleur varie de blanc (pas de fraude) à noire (toutes les transactions sont frauduleuses). On distingue clairement de petites communautés où il y a beaucoup de fraudes.

Ces communautés de marchands suspects sont ensuite visualisées : chaque nœud est maintenant un marchand, et la couleur (noir : fraude, blanc : pas de fraude) correspond au montant de la fraude dans la figure du milieu, ou au taux de fraude dans la figure de droite. Cette représentation permet aux équipes d'investigation de la fraude de focaliser leurs enquêtes sur des groupes de marchands à fort taux de fraude, très connectés entre eux et ayant eu des achats par un grand nombre de mêmes cartes (l'épaisseur du lien est le nombre de cartes communes). Ces semi-cliques de marchands fonctionnent comme des gangs organisés s'échangeant des lots de numéros de cartes dérobés.

Ces exemples montrent ainsi quelques usages intéressants et innovants qu'on peut faire de la visualisation des réseaux, dans des contextes très variés.

5. La personnalisation

Pourquoi personnaliser ?

Les systèmes de recommandation sont apparus dans les années 1990 pour aider les utilisateurs à trouver des produits qui les intéresseraient dans la masse de produits existants. Amazon a réussi à devenir le site que nous connaissons grâce à son algorithme de recommandation [18] et, depuis le prix de 1 million de dollars offert par Netflix en 2006 [8], la recherche sur les systèmes de recommandation est un domaine très actif. La recommandation est aujourd'hui une fonctionnalité critique pour les sites de e-commerce ou les sites sociaux.

Nous allons donner ici quelques exemples d'utilisation des réseaux sociaux pour produire des recommandations.

La recommandation d'amis

La recommandation d'amis est une fonctionnalité offerte par tous les sites sociaux (Facebook, LinkedIn...). Skyrock souhaitait ouvrir cette fonctionnalité à ses membres. Nous avons donc exploité le réseau d'amis décrit au paragraphe précédent. Pour recommander des amis à un utilisateur A, on s'appuie sur deux constatations :

(1) Les amis de mes amis peuvent sans doute être mes amis : on va donc rechercher, pour recommander des amis à A, les nœuds ayant le plus d'amis communs avec A. Cela revient à faire un appariement de graphes, c'est-à-dire à rechercher les nœuds B qui maximisent l'index de Jaccard (c'est-à-dire la similarité) entre les listes d'amis de A et de B :

$$Jaccard(A,b) = \frac{Nb\_Amis\_Communs(A,B)}{Deg_A + Deg_B - Nb\_Amis\_Communs(A,B)}.$$

(2) La *loi de localité* : quand deux nœuds sont reliés à A, il y a de fortes chances qu'ils soient reliés entre eux. Le coefficient de *clustering* ou coefficient d'agglomération mesure cette propriété :

$$Coeff\_Clust(A/B) = \frac{2(Nb\_Triang_A + Nb\_Amis\_Communs(A,B))}{Deg_A(Deg_A+1)}.$$

Quand on recommande les nœuds ayant le plus d'amis en commun, on essaie d'augmenter le coefficient de *clustering* du nœud.

On peut ainsi fournir, pour chaque nœud A, une liste triée de nœuds B d'amis recommandés en appliquant successivement les critères $Jaccard(A,B)$ puis $Coeff\_Clust(A/B)$.

Skyrock a mis en production cette technique en proposant chaque matin une liste personnalisée de 20 recommandations d'amis à ses membres. Les résultats ont été

8. http://www.netflixprize.com/

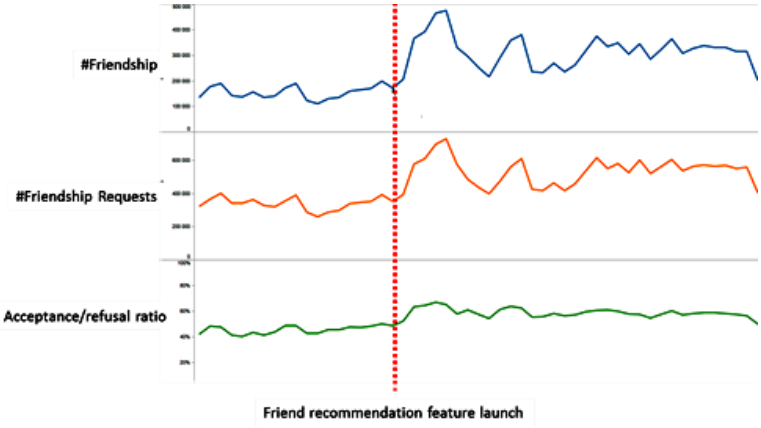

FIGURE 8. Recommandation d'amis sur le site Skyrock. En horizontal le nombre de jours, en vertical le nombre total d'amis (haut), de demandes d'amitié (milieu) et le rapport du nombre de demandes acceptées sur le nombre de refusées (bas).

immédiats comme le montre la figure 8 : le nombre de demandes de création d'amis et le nombre d'acceptations ont doublé, ces taux se sont maintenus par la suite.

La recommandation de contenus

Un site comme Skyrock contient de très nombreux contenus postés par les membres : les membres publient des contenus (textes, photos, vidéos) et des tags décrivant ces contenus. Nous pouvons donc là encore construire un graphe bipartite liant les *contenus* aux *tags*, puis le projeter en un graphe de Contenus, où deux contenus sont liés s'ils ont beaucoup de tags en commun.

Pour chaque page où se trouve un contenu, on peut alors proposer une liste des contenus connectés, dans le réseau projeté des contenus, à ceux de la page : c'est un exemple très simple de recommandation statique (analogue aux techniques de recommandation *content-based*), qui ne dépend pas directement des actions passées de l'utilisateur. Nous avons mis en place cette technique, en utilisant les traces des visites sur la plateforme Skyrock sur un mois, et observé une augmentation du taux de consultation des contenus. Cette augmentation est mesurée par l'indicateur MAP@5, qui indique la proportion de contenus consultés présents parmi les cinq que l'on recommande à l'utilisateur (Figure 9, gauche).

Une autre application des principes de recommandation personnalisée est le ciblage publicitaire. Il s'agit ici de choisir la bannière de publicité à montrer à un utilisateur. Skyrock utilisait l'outil de diffusion de bannières Open Ad Stream de `www.realmedia.com`. Nous avons construit le réseau bipartite liant les *utilisateurs*

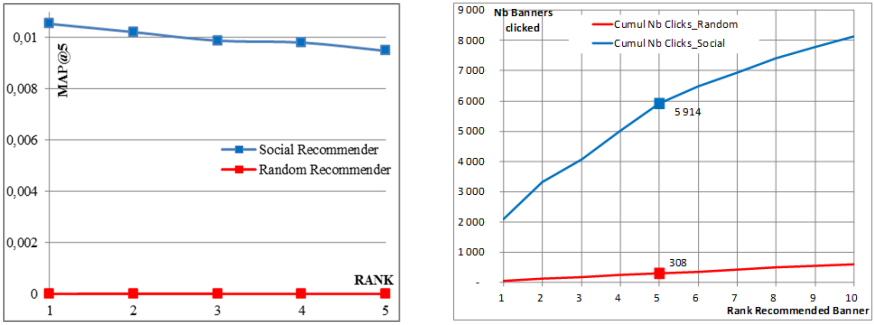

FIGURE 9. Recommandation de contenus (à gauche) et de bannières (à droite) sur Skyrock.

aux *bannières*, dans lequel un lien indique que l'utilisateur a cliqué sur la bannière. Ces données représentent un très gros volume : en septembre 2011, on a recueilli 8,3 milliards de lignes pour 2400 bannières. On utilise le graphe projeté des bannières, et on recommande à chaque utilisateur les bannières qui, dans ce réseau, sont connectées à celles sur lesquelles il a cliqué. Les résultats de ce modèle, calculés sur les deux premières semaines de septembre 2011 et appliqués sur les deux dernières semaines, fournissent une multiplication par 20 du nombre de clics, comme le montre la courbe de droite sur la figure 9.

La recommandation de parcours

Dans le cadre du projet AMMICO [10], nous avons réfléchi avec des musées français à des dispositifs pour aider les visiteurs des collections ou des expositions temporaires. Les besoins des visiteurs sont variés. Certains disposent de peu de temps et désirent voir quelques œuvres importantes, d'autres veulent aller plus loin et cherchent des informations ou des suggestions d'œuvres complémentaires, pendant ou après leur visite. Par ailleurs, les grands musées ont à gérer de nombreuses contraintes. Les plus fréquentés aimeraient mieux répartir le public, afin d'éviter l'effet « Joconde » (une salle trop pleine, des salles voisines presque vides alors qu'elles contiennent des œuvres de grand intérêt). Nous avons proposé, avec un industriel spécialiste du sujet, de mettre au point un audioguide moderne embarquant des fonctions innovantes : géolocalisation précise du visiteur, permettant d'enregistrer sa trajectoire dans le musée et de lui proposer des informations en lien avec sa position, module de recommandation couplé à un moteur de recherche contextuel, portail Web de post-visite offrant aux visiteurs de revoir les œuvres qu'ils ont découvertes au musée, d'en savoir plus, et de bénéficier de conseils personnalisés.

L'un des défis de ce type de projet est de préserver la vie privée des utilisateurs, alors que les services que l'on pourra leur proposer seront d'autant plus pertinents

que nous en saurons beaucoup sur eux : trajectoire dans le musée, traces d'utilisation du site Web, mais aussi informations extérieures que ces visiteurs pourraient avoir publiées sur des réseaux sociaux comme Twitter ou Facebook.

Nous avons développé des systèmes de recommandation exploitant à la fois la description des œuvres (catalogues des musées) et le comportement des utilisateurs [14]. Ces systèmes sont actuellement en test dans plusieurs musées parisiens.

La recommandation de recettes

Les sites Web culinaires sont parmi les plus fréquentés. Chaque jour, des millions de personnes y partagent des recettes, les annotent, les discutent. Plus de deux millions de recettes de cuisines différentes, originaires de très nombreux pays, sont ainsi accessibles sur un site comme `http://www.keyingredient.com/`. Ces sites sont avant tout des sites sociaux : leurs utilisateurs y accordent une grande importance aux avis et informations des autres.

Dans le cadre du projet *Open Food System*, nous avons travaillé avec les entreprises SEB et Cohéris au développement de méthodes d'analyse de données de grande taille et des réseaux sociaux, pour proposer différents services : détection de nouvelles thématiques (émergence de nouvelles modes culinaires), recommandation de recettes ou d'ingrédients. Dans le cadre de ce projet, d'autres partenaires s'intéressent à la prise en compte de contraintes alimentaires (produits disponibles, budget, allergies) ou d'objectifs diététiques (repas équilibrés, régimes).

Les systèmes de recommandation sociaux que nous avons ici mis en œuvre reposent sur l'analyse des communautés. On détecte des communautés d'utilisateurs partageant les mêmes goûts ou intérêts, et on recommande à chacun des recettes ayant plu aux membres de sa communauté (Figure 10).

6. L'utilisation pour les modèles prédictifs

La construction de modèles prédictifs par les algorithmes de *data mining* repose sur une hypothèse fondamentale : on suppose que les observations disponibles sont i.i.d., c'est-à-dire indépendantes, identiquement distribuées, issues du tirage selon une distribution fixe (inconnue). Comme nous l'avons vu dans les exemples précédents, beaucoup des entités qu'on observe ne sont en fait pas indépendantes, mais entretiennent des relations qu'on peut modéliser dans un réseau complexe. Nous allons montrer sur un exemple comment on peut exploiter une technique d'analyse de réseau social pour prendre en compte ces relations dans la fabrication du modèle prédictif.

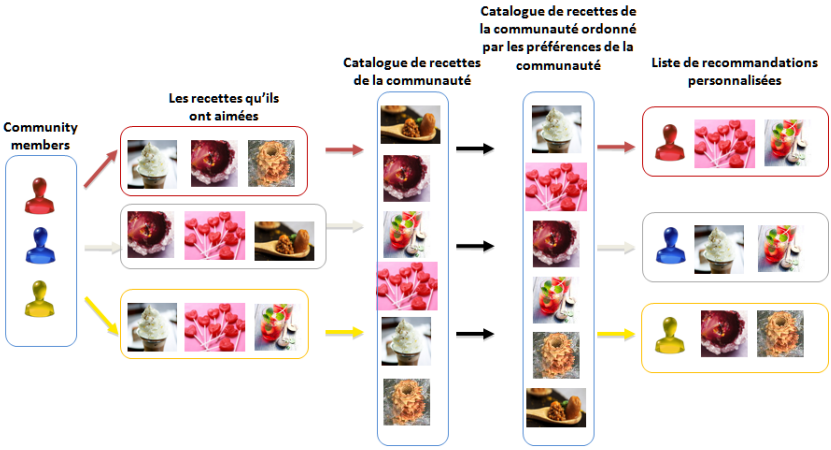

FIGURE 10. Recommandation de recettes.

La méthode

La méthode générale consiste à représenter les interactions entre les entités dans un réseau unipartite, ou obtenu par projection d'un réseau bipartite. Ensuite, des *variables sociales* sont extraites du réseau pour chacune des entités (les nœuds du réseau) : ces variables sont par exemple le degré, l'indice de la communauté, le coefficient de *clustering*, un attribut moyen sur les voisins du nœud, sur les membres de sa communauté... On augmente ainsi la variété des attributs du nœud en introduisant des variables qui représentent les interactions (faute d'un cadre théorique pour le *data mining* qui s'abstrairait de l'hypothèse i.i.d.). On sait, depuis les travaux sur le *Big Data* [6], que l'augmentation de la variété permet d'augmenter l'efficacité des modèles. C'est ce que nous allons illustrer ici.

L'exemple de la rétention

Skyrock sait que si un membre reste inactif pendant 120 jours, alors il ne reviendra probablement jamais. On veut donc essayer de prévoir les membres qui vont devenir inactifs pour éventuellement essayer de les retenir par des recommandations personnalisées, des applications gratuites...

Nous construisons donc le réseau d'amis qui a déjà été décrit plusieurs fois. Sur le mois de mars 2011, nous identifions les utilisateurs actifs sur le mois, puis nous cherchons ceux qui sont devenus inactifs le mois suivant. Nous entraînons un modèle prédictif sur ces données en utilisant un modèle de communauté locale décrit dans [21]. Les résultats obtenus (voir Figure 11) montrent que les variables sociales augmentent significativement (au vu de la taille des données) les performances du modèle, et que

Méthode	Nb moyen de nœuds utilisés	AUC[9]
Attributs du nœud & communauté locale	21	0,832
Attributs du nœud & second voisinage	71734	0,826
Attributs du nœud & premier voisinage	598	0,823
Attributs du nœud seul	1	0,815

FIGURE 11. Amélioration d'un modèle de rétention (*churn*) avec des variables sociales.

la notion de communauté locale permet de ne considérer qu'un nombre réduit de voisins dans le graphe social, ce qui permet, si on pré-calcule astucieusement les communautés, d'effectuer des prévisions en temps réel même dans de très grands réseaux assez denses.

L'exemple de la détection de la fraude

Les transactions sur Internet sont traitées par les réseaux de cartes bancaires (Visa, Mastercard...) pour autoriser ou refuser les transactions, mais aussi pour détecter les transactions frauduleuses. Quand il y a une suspicion de fraude sur une transaction, on met la carte en alerte et on contacte la banque et le porteur de la carte. Plus on détecte la fraude tôt, plus le coût de la fraude est réduit. Nous avons travaillé sur les données disponibles au GIE CB, dans le cadre du projet eFraudBox [8]. On dispose de l'intégralité des transactions réalisées sur Internet par les porteurs de carte Visa ou MasterCard, provenant de banques françaises. Ces données représentent 37 variables pour chaque transaction (numéro de carte, date d'expiration, banque émetteur, identifiant du marchand, SIRET, pays, activité du marchand, banque du marchand et pays de la banque, terminal utilisé, date de la transaction, montant...).

Un modèle simple (*baseline*), utilisant uniquement les 37 variables initiales décrivant les transactions fournit des performances très faibles : 8,2 % en précision et 1,4 % en rappel, où précision et rappel sont définis par :

$$Précision = \frac{VP}{A}, \qquad Rappel = \frac{VP}{F}.$$

Précision est le taux de cartes en alerte réellement frauduleuses (A est le nombre de cartes mises en alerte) et *Rappel* est le taux de fraude capturé par le modèle (F est le nombre total de cas de fraude). Les vrais positifs (VP) sont les cas de fraude que le modèle a identifiés comme tels. Les performances visées en opérationnel sont de 70 % en précision et 30 % en rappel. Pour obtenir ces performances, nous avons tout d'abord augmenté la variété des attributs sur chaque transaction :

— On calcule 300 agrégats cartes et 366 agrégats marchands ;

9. L'AUC (*Area Under the Curve*) est l'aire sous la courbe représentant le taux de vrais positifs en fonction du taux de faux positifs.

Variables	Nombre	Modèle	Rappel	Précision
Initiales (transactions)	37	Baseline	1,40 %	8,18 %
Agrégats Carte	300	Baseline + Agrégats	9,13 %	19,00 %
Agrégats Marchand	366			
Variables sociales Carte	195	Baseline + Agrégats + Agrégats sociaux	9,09 %	40,58 %
Variables sociales Marchand	99			
Total	**997**			

FIGURE 12. Détection de la fraude.

— On calcule 195 variables sociales cartes et 99 variables sociales marchands de la façon suivante : on construit le réseau bipartite Cartes × Marchands. On le projette en un réseau Cartes et un réseau Marchands. Puis pour chaque carte (resp. marchand), on calcule les variables sociales comme décrit précédemment dans le réseau Cartes (resp. Marchands).

On passe ainsi de 37 à 997 variables. Comme le montre la figure 12, l'ajout d'agrégats et de variables sociales permet à chaque fois de *doubler la précision*, sur un problème particulièrement difficile puisqu'il s'agit de détecter la fraude qui ne représente que 0,03 % des cas. Pour augmenter le rappel, il faut une technique de segmentation décrite dans [8] qui permet d'atteindre finalement 16,46 % en rappel et 60,89 % en précision.

Ces deux exemples démontrent l'apport des techniques d'analyse de réseaux sociaux pour l'amélioration des modèles prédictifs. Dans tous les cas où nous avons utilisé cette technique, les performances ont toujours été très largement augmentées : il n'existe guère de technique capable de doubler les performances d'un modèle !

7. Conclusion

Nous avons présenté ici quelques cas concrets d'utilisation des techniques de l'analyse des réseaux sociaux pour différentes applications dans lesquelles ils ont apporté des méthodes nouvelles et de nouveaux moyens de comprendre et prévoir : description, recommandations personnalisées et prévision.

Le domaine est très actif et il suffit de suivre les principales conférences du domaine, comme par exemple *International Conference on Advances in Social Networks Analysis and Mining* (ASONAM), *Recommender Systems Conference* (RecSys) ou *Conference on Knowledge Discovery and Data Mining* (KDD), pour s'en convaincre. Nous n'avons bien sûr couvert qu'une très faible partie du domaine. Le lecteur intéressé pourra se référer aux actes de ces conférences pour davantage d'exemples.

Notons que la communauté académique française compte quelques équipes spécialisées dans l'analyse des réseaux sociaux, comme le LABRI à Bordeaux avec son

logiciel de visualisation Tulip [10], l'équipe des réseaux complexes au LIP6 à Paris [11], la chaire réseaux sociaux à l'Institut Mines Telecom [12] et Telecom Bretagne [13], entre autres.

Remerciements

Nous remercions l'ANR, le programme FUI et les pôles de compétitivité qui ont financé les projets résumés ici : CADI (ANR : 2007-2010), CEDRES-ExDeuss (ANR-DGCIS : 2009-2013), eFraudBox (ANR : 2010-2013), AMMICO (FUI 13 : 2012-2015), OFS (PSPC : 2012-2016) et REQUEST (Big Data : 2014-2018).

Références

[1] C.C. Aggarwal, H. Wang. *Managing and Mining Graph Data*. Springer, 2010.

[2] S. Asur, B. A. Huberman. Predicting the Future With Social Media. Proceedings of the 2010 IEEE / WIC / ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 492–499, 2010.

[3] A.-L. Barabási. *Linked*. Plume, Penguin Group, 2002.

[4] A.-L. Barabási. *Network Science*. Cambridge University Press, à paraître, mai 2016.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte and E. Lefebvre. Fast unfolding of communities in large networks. *J. Stat. Mech.* P10008, volume 2008, October 2008.

[6] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), pp. 78–87, 2012.

[7] P. Erdös and A. Rényi. On random graphs. *Publicationes Mathematicae*, 6, pp. 290–297, 1959.

[8] F. Fogelman Soulié, A. Mekki, S. Sean, P. Stepniewski. Utilisation des réseaux sociaux dans la lutte contre la fraude à la carte bancaire sur Internet. In *Apprentissage Artificiel & Fouille de Données*. Y. Bennani, E. Viennet eds, Revue des Nouvelles Technologies de l'Information, RNTI-A-6, Hermann, pp. 99–119, 2014.

[9] S. Fortunato. Community detection in graphs. *Physics Reports*, vol. 486, Issues 3–5, pp. 75–174, February 2010. http://www.sciencedirect.com/science/journal/03701573

[10] R. Fournier, E. Viennet, S. Sean, F. Fogelman Soulié, M. Bénaïche. AMMICO : social recommendation for museums. *Digital Intelligence – DI2014*, Nantes, France, September 17–19, 2014.

[11] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, June 11, 2002.

[12] B.A. Huberman. *The laws of the Web. Patterns in the Ecology of Information.* MIT Press, Cambridge, 2001.

[13] B.J. Jansen, M. Zhang, K. Sobel, A. Chowdury. Twitter power : Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, Vol. 60, Issue 11, pp. 2169–2188, November 2009.

10. http://www.labri.fr/
11. http://www.lip6.fr/recherche/team.php?id=790
12. https://chairereseaux.wp.mines-telecom.fr/en/
13. http://perso.telecom-bretagne.eu/cecilebothorel/

[14] I. Keller, E. Viennet. Recommender Systems for Museums – Evaluations on a Real Dataset. Fifth International Conference on Advances in Information Mining and Management, July 2015.

[15] J.M. Kleinberg, S. Lawrence. The structure of the Web. *Science* (Washington), Vol. 294, no. 5548, pp. 1849–50, November 30, 2001.

[16] J. Leskovec, L. A. Adamic and B. A. Huberman. The dynamics of viral marketing. Proceedings of the 7th ACM Conference on Electronic Commerce. *ACM Transactions on the Web (TWEB)*, Vol. 1, Issue 1, Article No. 5, May 2007.

[17] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, N. Glance. Cost-effective Outbreak Detection in Networks. Proceedings of the 13rd ACM SIGKDD Conference on Knowledge Discovery and data mining KDD'07, San Jose, USA, pp. 420–429, 2007.

[18] G. Linden, B. Smith and J. York. Amazon.com Recommendations Item-to-Item Collaborative Filtering. *IEEE Internet Computing*, 7(1), pp. 76–80, 2003.

[19] S. Milgram. The small world problem. *Psychology Today*, 1, p. 61, 1967.

[20] N. Memon, D. L. Hicks. Detecting Core Members in Terrorist Networks : a Case Study. In *Mining massive Data Sets for Security : Advances in data mining, search, social networks and text mining and their applications to security*. F. Fogelman-Soulié, D. Perrotta, J. Pikorski, R. Steinberger eds., IOS Press, pp. 345–356, 2008.

[21] B. Ngonmang, E. Viennet, S. Sean, P. Stepniewski, F. Fogelman-Soulié, R. Kirche. Monetization and Services on a Real Online Social Network Using Social Network Analysis. ICDM'2013, 2013 IEEE 13th International Conference on Data Mining. `http://www.dataminingcasestudies.com`, Fifth Workshop on Data Mining Case Studies and Practice Prize (DMCS-5), pp. 185–193, Dallas, Texas, December 7–10, 2013.

[22] S. Wasserman, K. Faust. *Social network analysis : Methods and applications*. Cambridge University Press, 1994.