Les capitalistes sociaux sur Twitter : détection via des mesures de similarité

Nicolas Dugué*, Anthony Perez*

* LIFO - Université d'Orléans

Résumé. Les réseaux sociaux tels que Twitter font partie du phénomène de *Déluge des données*, expression utilisée pour décrire l'apparition de données de plus en plus volumineuses et complexes. Pour représenter ces réseaux, des graphes orientés sont souvent utilisés. Dans cet article, nous nous focalisons sur deux aspects de l'analyse du réseau social de Twitter. En premier lieu, notre but est de trouver une méthode efficace et haut niveau pour stocker et manipuler le graphe du réseau social en utilisant des ressources informatiques raisonnables. Cet axe de recherche constitue un enjeu majeur puisqu'il est ainsi possible de traiter des graphes à échelle réelle sur des machines potentiellement accessibles par tous. Ensuite, nous étudions les *capitalistes sociaux*, un type particulier d'utilisateurs de Twitter observé par Ghosh et al. (2012). Nous proposons une méthode pour détecter et classifier efficacement ces utilisateurs.

1 Introduction

Contexte. Depuis plusieurs années, dans les secteurs de l'Internet, de l'analyse décisionnelle ou encore de la génétique sont collectées et analysées des données de plus en plus volumineuses et complexes. Ce phénomène connu sous le nom de *Déluge des données* (ou *Big Data*) soulève de nombreuses problématiques. En particulier, être capable de stocker, partager et analyser de telles quantités de données constitue un enjeu d'étude essentiel, comme le soulignent Schuett et Pierre (2012). La théorie des graphes est particulièrement appropriée pour étudier les réseaux sociaux, où les connexions entre utilisateurs peuvent facilement être représentées et analysées en utilisant des graphes, le plus souvent orientés. Dans cet article, nous considérons le *graphe des relations entre utilisateurs (anonymisés) de Twitter* créé en 2009 par Cha et al. (2010). Ce graphe orienté contient plus de 50 millions de sommets et près de 2 milliards d'arcs.

Les capitalistes sociaux. Nous nous intéressons particulièrement au comportement d'utilisateurs particuliers nommés *capitalistes sociaux*, observé par Ghosh et al. (2012). Ces utilisateurs, qui ne sont ni des spammeurs, ni des robots, partagent un objectif commun : acquérir un nombre maximum d'utilisateurs qui les suivent -*followers*. En effet, plus le nombre de *followers* d'un utilisateur est élevé, plus il peut être influent sur le réseau. Au-delà de cet intérêt évident, le nombre de *followers* a un impact direct sur le classement des *tweets* de l'utilisateur sur le moteur de recherche de Twitter. Ces utilisateurs ne sont pas sains pour un réseau social : en suivant des utilisateurs sans regarder le contenu de leurs tweets, les capitalistes sociaux donnent de l'influence à des utilisateurs tels que les spammeurs.

Notre contribution. Les résultats que nous proposons dans cet article sont de deux types. Nous nous concentrons tout d'abord sur la recherche de méthodes *efficaces* et *haut niveau* permettant de stocker et manipuler le graphe des relations entre utilisateurs de Twitter en ayant recours à des ressources informatiques raisonnables [1]. Nous nous intéressons ensuite à la détection des capitalistes sociaux. En particulier, nous montrons que ces derniers peuvent être détectés efficacement en appliquant des *mesures de similarité* sur les voisinages du graphe des relations. Pour notre étude, nous utilisons un graphe collecté en 2009 par Cha et al. (2010), contenant les utilisateurs *anonymisés* de Twitter ainsi que les relations qui existent entre eux. Plus précisément, pour mettre en oeuvre nos mesures, nous considérons le *graphe des spammeurs* de Twitter, qui contient près de 40000 spammeurs (détectés par Ghosh et al. (2012)) ainsi que leurs voisins, pour un total de 15 millions de sommets et 1 milliard d'arcs (Section 2). Nous verrons dans la suite de cet article que travailler sur un tel graphe est suffisant pour nos besoins. Finalement, pour valider notre méthode de détection, nous comparons nos résultats à une liste de 100000 capitalistes sociaux potentiels détectés de manière *ad-hoc* par Ghosh et al. (2012) lors de leur étude sur les spammeurs. Nous observons que nos algorithmes détectent une grande majorité de ces utilisateurs, qui ont la quasi-totalité de leur voisinage inclus dans le graphe des spammeurs (Section 3).

2 Graphe des spammeurs : stockage et définition

Stockage. Dans leurs travaux respectifs, Cha et al. (2010) et Ghosh et al. (2012) ne donnent aucun détail sur la méthode qu'ils ont employée pour manipuler le graphe de Twitter. Dans cet article, notre premier objectif est ainsi de trouver un procédé haut niveau pour stocker et traiter le graphe des relations de Twitter en utilisant des ressources informatiques raisonnables. La méthode que nous suggérons dans cet article peut donc être reproduite facilement sur un simple serveur avec un unique processeur disposant d'une certaine quantité de mémoire vive (environ 24 Go pour étudier efficacement le graphe des spammeurs). Remarquons que les mesures développées dans l'article auraient pu être réalisées via des traitements sur la liste d'adjacence du graphe mais cela n'est pas une méthode haut niveau extensible à d'autres problèmes.

Pour parvenir à cet objectif, nous avons utilisé des bases de données pour stocker le graphe et l'analyser. Nous avons exploré la possibilité d'utiliser tout type de bases de données, qu'elles soient dites SQL, NoSQL ou orientées graphes. Les résultats de nos expérimentations montrent que, même pour de simples mesures telles que celles que nous mettons en oeuvre, il est plus efficace d'utiliser une base de données orientée graphes. Par exemple, si MySQL permet de rapidement charger les données, l'exécution de requêtes calculant l'intersection de deux voisinages (un outil nécessaire pour nos mesures) est relativement lente (plusieurs jours pour les exécuter sur tous les sommets ou plusieurs jours pour poser les index efficaces). Avec Cassandra (NoSQL), le simple fait d'obtenir le degré de chaque sommet peut demander de nombreuses heures de traitement. Par conséquent, nous avons essayé plusieurs bases de données orientées graphes, telles que OrientDB ou Neo4j. Dans les deux cas, nous n'avons pas été capables de charger le graphe en un temps raisonnable (moins d'une semaine). Finalement, **Dex** (voir Martínez-Bazan et al. (2012)) est apparue comme une solution viable pour plusieurs

1. Pour nos études, nous avons travaillé avec une seule machine : AMD Opteron(tm) Processor 6174 800 Mhz 12 cores (nos algorithmes ne sont pas parallélisés), avec 64 Go RAM.

raisons : orientée graphes, haute performance et dotée d'une API haut niveau. Sur simple demande, nous avons obtenu une licence temporaire fournissant un accès complet à toutes les fonctionnalités. Cependant, si nous avons été capables de stocker le graphe des relations de Twitter en utilisant Dex (avec un temps de chargement de quelques heures), nos algorithmes n'ont parfois pas pu s'exécuter sur la totalité du graphe à cause de dysfonctionnements. Ces problèmes techniques -détectés par nos expérimentations- sont toujours en cours de correction. Cependant, ils ne sont pas apparus sur un graphe de plus petite taille (dit *des spammeurs*) qui reste cohérent avec nos objectifs, et nos algorithmes s'exécutent ainsi en quelques heures.

Définition. En nous servant du *graphe des relations entre utilisateurs de Twitter* $D = (V, A)$, où V représente l'ensemble des utilisateurs (*anonymisés*) et A l'ensemble des *arcs*, dirigés de u vers v si l'utilisateur u *suit* (ou *follow*) l'utilisateur v, nous avons calculé le *graphe des spammeurs* $S = (V', A')$. Ici, V' représente un ensemble de spammeurs et leurs voisins, et A' les arcs existant entre ces sommets. Pour obtenir un tel graphe, nous utilisons une liste de **40000 spammeurs** (utilisateurs suspendus par Twitter pour diffusion de liens dangereux) fournie par Ghosh et al. (2012). En faisant cela, nous préservons **15 millions** de sommets et un peu plus d'**1 milliard** d'arcs. Cela représente environ la moitié du graphe des relations entre utilisateurs de Twitter fourni par Cha et al. (2010). Etant donné $v \in V'$, nous définissons $N^+(v)$ (resp. $N^-(v)$) comme l'ensemble des voisins *sortants* (resp. *entrants*) de v. Par conséquent, le degré *sortant* (resp. *entrant*) d'un sommet est donné par $|N^+(v)|$ (resp. $|N^-(v)|$). Observons que le degré entrant d'un utilisateur correspondant aux nombres de personnes qui le suivent (*followers*), alors que son degré sortant reprśente le nombre de personnes qu'il suit.

Nous allons voir que détecter des utilisateurs particuliers dans ce graphe avec des mesures de similarité sur les voisinages peut permettre d'obtenir des informations pertinentes dans le graphe des relations. En effet, la majorité des utilisateurs détectés par nos algorithmes ont une grande proportion de leur voisinage dans le graphe des spammeurs.

3 Capitalistes sociaux

De façon identique aux comportements observés sur Internet, où les administrateurs de sites webs effectuent de l'*échange de liens* dans le but d'accroître leur visibilité, certains utilisateurs cherchent à obtenir un maximum de *followers* afin d'augmenter leur influence. Pour parvenir à cet objectif, ces utilisateurs exploitent deux techniques relativement simples et basées sur la réciprocation du lien *follow* : **FMIFY** (Follow Me and I Follow You -l'utilisateur assure à ses followers qu'il les suivra en retour) et **IFYFM** (I Follow You, Follow Me -ces utilisateurs suivent d'autres utilisateurs en espérant que ceux-ci les suivent en retour). Ces comportements ont été mis en lumière par Ghosh et al. (2012) : lors d'une étude sur les *spammeurs*, ils ont observé une classe d'utilisateurs *réels* (*i.e.* ni des spammeurs, ni des faux comptes) répondant beaucoup aux sollicitations des spammeurs, les qualifiant ainsi de capitalistes sociaux.

Définition 1 (Capitaliste social)**.** *Un* capitaliste social *est un utilisateur appliquant les principes FMIFY/IFYFM dans le but d'augmenter son influence sur le réseau social Twitter.*

Il est intéressant de noter que plusieurs comptes célèbres sur Twitter (comme par exemple celui de **Barack Obama**) sont connus pour avoir appliqué ces principes.

Mesures de similarité. Dans le but de détecter des capitalistes sociaux, nous utilisons deux mesures de similarité sur le voisinage des utilisateurs, à savoir l'*indice de chevauchement* (introduit par Simpson (1943)) et le *ratio*. La première nous permet de détecter de potentiels capitalistes sociaux, alors que la dernière nous sert à classifier ces capitalistes sociaux selon leur utilisation de l'un ou l'autre des principes **FMIFY** et **IFYFM**.

Définition 2 (Indice de chevauchement)**.** *Soient deux ensembles A et B, l'*indice de chevauchement *de A et B (qui a valeur dans $[0..1]$) est donné par $C(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$.*

Pour chaque sommet v, nous appliquons l'indice de chevauchement sur les ensembles $N^+(v)$ et $N^-(v)$. Cela nous permet de détecter des utilisateurs susceptibles d'être des capitalistes sociaux. En effet, les voisinages entrants et sortants de ces derniers doivent être fortement liés ; en d'autres termes, ils doivent suivre la majorité de leurs followers (principe **FMIFY**), ou inversement être suivis par la majorité des utilisateurs qu'ils suivent (principe **IFYFM**). En particulier, cela signifie que leur indice de chevauchement doit être proche de 1, l'ensemble $N^+(v) \cap N^-(v)$ étant quasiment inclus dans $N^-(v)$ ou $N^+(v)$, respectivement. Nous utilisons par la suite la Définition 3 pour classifier ces utilisateurs.

Définition 3 (Ratio)**.** *Etant donné un sommet v, le* ratio *de v est obtenu par $R(v) = \frac{|N^+(v)|}{|N^-(v)|}$.*

Intuitivement, les utilisateurs qui suivent le principe **IFYFM** doivent avoir un ratio supérieur à 1, tandis que les utilisateurs qui appliquent **FMIFY** doivent avoir un ratio inférieur à 1. Dans les deux cas, le ratio attendu doit être proche de 1. Nous observons cependant un comportement qui engendre un ratio très inférieur à 1 chez certains utilisateurs, que nous appelons *passifs*. Contrairement aux autres capitalistes sociaux, ces utilisateurs considèrent leur degré entrant comme suffisant. Ils cessent donc d'utiliser les principes mentionnés ci-dessus mais continuent à accumuler des followers, notamment grâce à l'influence dont ils disposent.

Détection dans le graphe des spammeurs. Nous prétendons que détecter les capitalistes sociaux dans un tel graphe peut fournir des informations pertinentes à propos des capitalistes sociaux dans le graphe des relations entre utilisateurs. Pour illustrer cela, nous utilisons une liste de 100000 utilisateurs considérés comme des capitalistes sociaux par Ghosh et al. (2012). Ces derniers ont été répérés de manière ad-hoc : en partant d'une liste de 40000 spammeurs potentiels, ils ont considéré comme capitalistes sociaux les utilisateurs répondant le plus aux sollicitations des spammeurs. Ces derniers doivent donc se retrouver dans le graphe des spammeurs. Plus précisément, comme l'illustre la Figure 1, la majorité de ces utilisateurs possède la quasi-totalité de leur voisinage dans le graphe des spammeurs. Ces premières observations nous permettent de valider la pertinence de nos mesures de similarité basées sur les voisinages, et donc de poursuivre l'étude des capitalistes sociaux sur le graphe des spammeurs.

%	entrants	sortants	entrants et sortants
90	88407	73431	72368
80	95107	90359	83860
70	96901	96047	89436

FIG. 1 – *Nombre de capitalistes sociaux parmi les* 100000 *détectés par Ghosh et al. (2012) ayant* % *de leurs voisinages (entrants, sortants) dans le graphe des spammeurs.*

Résultats. Nous présentons maintenant les résultats expérimentaux (voir à gauche de la Figure 2) que nous obtenons sur le graphe des spammeurs. La colonne $détectés$ présente le nombre de capitalistes sociaux potentiels observé, relativement aux contraintes posées sur le degré entrant (colonne $entrant$) et sur l'indice de chevauchement (qui est dans tous les cas supérieur à 0.8). Nous considérons ici uniquement des sommets ayant un degré entrant élevé, correspondant à des utilisateurs ayant utilisé les principes **FMIFY/IFYFM** avec succès.

entrant	ratio	détectés
> 500		137267
	> 1	86594
	$[0.7; 1]$	33291
> 10000		5344
	> 1	3378
	$[0.7; 1]$	1232
	< 0.7	734

entrant	ratio	liste Ghosh et al. (2012)	%
> 500		67523	67
	> 1	54870	55
	$[0.7; 1]$	12347	13
> 10000		4527	4
	> 1	3260	3
	$[0.7; 1]$	1188	1
	< 0.7	79	0

FIG. 2 – *Sur la gauche, les utilisateurs détectés par nos algorithmes. Sur la droite, le pourcentage d'utilisateurs de la liste de Ghosh et al. (2012) détectés par nos algorithmes.*

Quelle que soit la contrainte imposée sur le degré entrant, nous observons les comportements **IFYFM** et **FMIFY** : pour un degré supérieur à 500, 62% des utilisateurs ont un ratio supérieur à 1 et 24% ont un ratio entre 0.7 et 1, respectivement. Nous remarquons également que des capitalistes sociaux dits *passifs* sont présents lorsque le degré est supérieur à 10000 : 14% des 5344 sommets ayant un indice de chevauchement supérieur à 0.8 ont un ratio inférieur à 0.7.

Validation des résultats. Afin de confirmer nos observations, nous utilisons la liste de cent mille capitalistes sociaux détectés de manière ad-hoc par Ghosh et al. (2012). Rappelons que ces derniers devraient être détectés par notre méthode, un fait visible sur la table droite de la Figure 2. Nous aimerions mentionner qu'environ 12500 utilisateurs de la liste de Ghosh et al. (2012) possèdent moins de 500 followers. Pour mieux illustrer la cohérence de nos résultats, nous montrons maintenant que les utilisateurs détectés comme capitalistes sociaux potentiels par nos algorithmes ont leur voisinage presque entièrement contenu dans le graphe des spammeurs (Figure 3). Ainsi, la plupart des utilisateurs que nous détectons peuvent être considérés comme des capitalistes sociaux dans le graphe des relations. En effet, par la Définition 2, tout utilisateur ayant moins de 10% de son voisinage à l'extérieur du graphe des spammeurs aura un indice de chevauchement d'*au moins* 0.72 dans le graphe des relations.

entrant	ratio	détectés	90%-entrants	90%-sortants	90%-entrants et sortants
> 500	> 1	86594	71984	60690	59585
> 500	$[0.7; 1]$	33291	22331	23288	21691
> 10000	> 1	3378	3268	3219	3217
> 10000	$[0.7; 1]$	1232	1058	1044	1037
> 10000	< 0.7	734	45	200	37

FIG. 3 – *Capitalistes sociaux contenant* 90% *de leurs voisins dans le graphe des spammeurs.*

4 Perspectives de recherche

De nombreuses pistes relatives à l'étude des capitalistes sociaux restent à explorer. Une suite logique de notre travail consiste à créer plusieurs comptes Twitter appliquant les principes utilisés par les capitalistes sociaux et ainsi valider et mieux comprendre l'efficacité de ces stratégies. De plus, nous souhaitons étudier la position de ces utilisateurs dans les communautés du graphe de Twitter. En effet, les capitalistes sociaux sont très fortement liés avec leur voisinage et sont souvent des sommets de degré élevé. Intuitivement, à cause de ces caractéristiques, ils doivent avoir un rôle important au sein ou entre les communautés. Nous voudrions ainsi voir la place de ces utilisateurs dans des communautés détectées avec des méthodes telles que celle proposée par Blondel et al. (2008).

Références

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journ. of Stat. Mechanics : Theory and Experiment 2008*(10), P10008.

Cha, M., H. Haddadi, F. Benevenuto, et K. P. Gummadi (2010). Measuring User Influence in Twitter : The Million Follower Fallacy. In *ICWSM*.

Ghosh, S., B. Viswanath, F. Kooti, N. K. Sharma, K. Gautam, F. Benevenuto, N. Ganguly, et K. Gummadi (2012). Understanding and Combating Link Farming in the Twitter Social Network. In *WWW*.

Martínez-Bazan, N., V. Muntés-Mulero, S. Gómez-Villamor, M. Águila-Lorente, D. Dominguez-Sal, et J.-L. Larriba-Pey (2012). Efficient Graph Management Based On Bitmap Indices. In *IDEAS*.

Schuett, T. et G. Pierre (2012). ConpaaS, an integrated cloud environment for big data. *ERCIM News 2012*(89).

Simpson, G. G. (1943). Mammals and the nature of continents. *American Journal of Science* (241), 1–41.

Summary

Social networks such as Twitter are part of the phenomenon called *Big Data*, a term used to describe very large and complex data sets. To represent these networks, oriented graphs are often used, containing dozens of millions of vertices and billions of arcs. In this paper, we are mainly focused on two different aspects of social network analysis. First, our goal is to find an high-level and efficient way to store and process such a social network graph, using reasonable computing resources (processor and memory). This constitutes a major issue because it allows to handle real-scale graph without using heavy machinery. Next, we turn our attention to the study of *social capitalists*, a specific kind of users on Twitter introduced by Ghosh et al. (2012). We suggest a method to detect and classify efficiently such users.