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

Nicolas Dugué, Anthony Perez

LIFO - Université d'Orléans
rue Léonard de Vinci B.P. 6759
F-45067 ORLEANS Cedex 2 FRANCE

RÉSUMÉ. *Les réseaux sociaux tels que Twitter et Facebook sont partie prenante du phénomène de déluge des données. Les graphes modélisant leurs utilisateurs et les liens existant entre eux représentent des dizaines de millions de sommets et plusieurs milliards d'arcs. Les traiter efficacement pour en analyser la topologie reste un challenge actuellement. Dans cet article, nous proposons une solution pour traiter de tels graphes, et nous intéressons également à la détection et à l'analyse d'une communauté particulière d'utilisateurs de Twitter appelés capitalistes sociaux.*

ABSTRACT. *Social networks such as Twitter or Facebook are part of the phenomenon called* Big Data*. Graphs modelizing their users and the links between them represent dozens of millions of vertices and billions of arcs. Being able to consider them efficiently in order to analyse their topology constitutes a major challenge. In this extented abstract, we propose a solution to deal with such graphs, and we study the detection and the characterization of a small community of users of Twitter, called* social capitalists*.*

MOTS-CLÉS : *réseau social, twitter, capitalistes sociaux, détection*

KEYWORDS: *social network, twitter, social capitalists, detection*

1. Introduction

1.1. *Contexte*

Depuis quelques années, dans les secteurs de l'Internet, de la business intelligence ou encore de la génétique sont collectées des données de plus en plus volumineuses et complexes (*Big Data*). Dans de nombreux cas, ces données peuvent être modélisées sous forme de graphes. L'explosion du nombre d'utilisateurs des réseaux sociaux engendre par exemple des graphes de plus en plus conséquents, dont il est bien souvent difficile d'analyser les propriétés structurelles. Dans le cadre de cet article, nous considérons le *graphe des relations entre utilisateurs de Twitter*. Plus particulièrement, nous nous intéressons aux comportements d'utilisateurs appelés *capitalistes sociaux*, dont l'objectif est d'obtenir un maximum de **followers** sur Twitter. En effet, plus le nombre de followers d'un utilisateur est grand, plus il est influent. Or l'influence, en plus d'être perçue par les autres utilisateurs, a un effet sur le classement des tweets de l'utilisateur sur le moteur de recherche de Twitter. Les capitalistes sociaux ont été mis en avant par [GHO 12], dans un article étudiant le comportement des **spammers**, utilisateurs suspendus par Twitter qui diffusaient des liens dangereux.

1.2. *Nos travaux*

Nous conjecturons qu'il est possible de détecter les capitalistes sociaux en utilisant de simples **mesures de similarité**. Afin de réaliser notre étude, nous utilisons un graphe collecté en **2009** (et mis à disposition par [CHA 10]). Plus précisément, afin de valider nos mesures de similarité, nous considérons le *graphe des spammers de Twitter*, comprenant $\mathbf{40000}$ spammers détectés par [CHA 10] ainsi que tous leurs voisins. Dans la mesure où les utilisateurs répondant aux spammers sont majoritairement des capitalistes sociaux, nous capturons ces derniers dans ce graphe, ce qui nous permet donc de réaliser des mesures cohérentes. Nous présentons tout d'abord la solution que nous avons choisie afin de stocker et d'analyser ce graphe (Section 2). Notre objectif était avant tout de considérer des **bases de données orientées graphe**. Par la suite, nous caractérisons le comportement de ces utilisateurs et proposons des méthodes pour détecter la communauté qu'ils forment (Section 3), avant de finalement présenter ces méthodes sur le graphe des spammers de Twitter. Finalement, nous concluons par quelques perspectives de recherche (Section 4).

2. Graphe des spammers : définition et stockage

2.1. *Notations*

A partir du *graphe des relations de Twitter* $D = (V, A)$, où V représente l'ensemble des utilisateurs et $uv \in A$, $\{u, v\} \in V$ si et seulement si l'utilisateur u est un *follower* de l'utilisateur v (*i.e.* u est abonné aux tweets de v), nous calculons dans un

premier temps le *graphe des spammers* $S = (V', A')$. Ici, V' représente les spammers et leurs voisins, et A' tous les arcs existant entre ces sommets. En partant d'un peu moins de **40000 spammers** fournis par [GHO 12], nous conservons ainsi **15 millions** de sommets et un peu plus d'**1 milliard** d'arcs, soit la **moitié** du graphe des relations de Twitter mis à disposition par [CHA 10]. Pour $v \in V'$, nous définissons $N^+(v)$ (resp. $N^-(v)$) comme l'ensemble des voisins *sortants* (resp. *entrants*) de v.

2.2. *Stockage*

Le principal problème est de stocker le graphe afin de pouvoir en analyser la structure efficacement. Dans leur article, Cha et al. [CHA 10] ne décrivent pas leur méthode de stockage. Cela constitue pourtant un enjeu fondamental pour l'analyse de grands graphes. Nous avons ainsi étudié de nombreuses bases de données, principalement orientées graphes (*e.g.* OrientDB, Neo4j). Pouvoir stocker le graphe dans une de ces structures était l'un de nos principaux objectifs. **Dex** ([MAR 12]) est apparue comme une solution viable pour de nombreuses raisons : orientée graphe et hautes performances et dotée d'une API haut niveau (voir Algorithme 1), riche et bien documentée. Sur simple demande, nous avons obtenu une licence temporaire nous permettant d'exploiter les capacités de Dex sans limite. Nous avons ainsi pu charger le graphe complet en quelques heures en utilisant une machine disposant de suffisamment de mémoire pour le contenir entièrement (1 processeur *AMD Opteron(tm) Processor 6174 800 Mhz 12 coeurs*, avec 64 Go de RAM). Néanmoins, des problèmes techniques liés à la gestion des très grands graphes par Dex ne nous ont pas permis de détecter les capitalistes sociaux sur ce graphe. Par ailleurs, ces problèmes techniques que nos expérimentations ont permis de détecter sont toujours en cours de correction. Ceci nous a donc poussé à travailler uniquement sur le *graphe des spammers de Twitter*. Sur ce graphe, nous avons pu par l'intermédiaire des mesures de similarité détecter les utilisateurs considérés comme étant des capitalistes sociaux en moins de trois heures.

3. Les capitalistes sociaux

3.1. *Définition*

A l'instar du web où les administrateurs de sites Internet effectuent de l'*échange de liens* de façon à accroître leur visibilité, les *capitalistes sociaux* cherchent à obtenir un maximum de *followers* afin d'accroître leur influence. Pour parvenir à leurs fins, nous observons que ces derniers adoptent deux techniques relativement simples, basées sur la réciprocation du lien de "follow" : (i) **FMIFY** (Follow Me and I Follow You - l'utilisateur promet aux utilisateurs les suivant de les suivre en retour) ; (ii) **IFYFM** (I Follow You, Follow Me - à l'inverse, ces utilisateurs suivent d'autres utilisateurs dans l'espoir d'être suivis en retour).

Les capitalistes sociaux ont été mis en avant par Ghosh et al. [GHO 12], lors d'une étude réalisée sur les *spammers* de Twitter. Ces derniers sont engagés dans un principe dit de *farm linking*, consistant à se faire suivre par un maximum d'utilisateurs.Ghosh et al. [GHO 12] se sont ainsi aperçus que les principaux utilisateurs répondant aux demandes de connexion des spammers sont des capitalistes sociaux. Ce résultat s'explique : comme décrit précédemment, les capitalistes sociaux utilisent les principes **FMIFY/IFYFM**, et suivent donc en retour les utilisateurs qui les suivent sans regard sur le contenu de leurs tweets.

3.2. *Mesures de similarité*

Les comportements décrits Section 3.1 amènent naturellement à penser qu'une corrélation forte existe entre les voisinages entrants et sortants des capitalistes sociaux. Ainsi, les utilisateurs respectant les principes **FMIFY/IFYFM** (les capitalistes sociaux que l'on dit *actifs*) auront une intersection de leurs voisinages entrants et sortants élevée. Nous verrons cependant que certains capitalistes sociaux (dits *passifs*) ont appliqué ces principes puis cessé après avoir obtenu une influence suffisante. Ainsi, le degré entrant de ces utilisateurs ayant acquis beaucoup de visibilité continue de croître alors que leur degré sortant reste stable. Par conséquent, si l'intersection de leurs voisinages sera faible, il existera une forte corrélation entre cette intersection et le voisinage sortant de ces utilisateurs. Pour détecter les capitalistes sociaux, qu'ils soient *actifs* ou *passifs*, nous introduisons une première mesure de similarité :

Definition 1. *Etant donnés deux ensembles A et B, l'*indice de chevauchement *de A et B (compris entre 0 et 1) est donné par la formule suivante :*

$$C(A,B) = \frac{|A \cap B|}{\min\left(|A|, |B|\right)}$$

Appliqué sur les ensembles $N^+(v)$ et $N^-(v)$ pour tout sommet $v \in V$, cet indice nous permet dans un premier temps de détecter les utilisateurs susceptibles d'être des capitalistes sociaux. Ayant appliqués les principes **FMIFY/IFYFM**, ces derniers auront en effet un indice de chevauchement proche de 1. Par la suite, afin de détecter si ces derniers sont *actifs* ou *passifs*, nous utilisons la notion de *ratio*. Le calcul de ce dernier est illustré Algorithme 1.

Definition 2. *Etant donné un sommet v, le* ratio *de v est défini par :*

$$R(v) = \frac{|N^+(v)|}{|N^-(v)|}$$

Un utilisateur ayant un indice de chevauchement et un ratio proches de 1 sera ainsi un capitaliste social *actif*. En revanche, un utilisateur avec un indice de chevauchement proche de 1 et un faible ratio sera considéré comme *passif*.

Algorithme 1: Extrait du code permettant de calculer les mesures de similatiré.

```
// Connexion à la base de données
DexConfig cfg = new DexConfig();
Dex d = new Dex(cfg);
Database db = d.open(nomBaseDeDonnees, true);
Session session = db.newSession();
Graph g = session.getGraph();
// Obtention des entiers décrivant les types de sommets et d'arcs du graphe
int user = g.findType("user") ; int edge = g.findType("follows");
// Récupération de la liste des sommets
Objects objs = g.select(user);
ObjectsIterator it = objs.iterator();
long v, in, out ; double ratio;
// Parcours de tous les sommets via itérateur
tant que it.hasNext() faire
    v = it.next();
    // Obtention des degrés sortant et entrant du sommet v
    out = g.degree(v, edge, EdgesDirection.Outgoing);
    in = g.degree(v, edge, EdgesDirection.Ingoing);
    // Obtention du ratio pour le sommet v
    si in ! = 0 alors
        ratio = new Double(out)/ new Double(in) ;
// La fermeture des itérateurs et objets de gestion de la base est impérative
it.close() ; objs.close() ; session.close() ; db.close() ; d.close();
```

Nous présentons maintenant les résultats que nous avons observés sur le graphe des spammers (Figure 1).

in	overlap	|sommets|	ratio	%
> 2000	> 0.8	40725		
> 2000	> 0.8		> 1	61
> 2000	> 0.8		$[0.7; 1]$	28
> 10000	> 0.8	5344		
> 10000	> 0.8		> 1	63
> 10000	> 0.8		$[0.7; 1.0]$	23
> 10000	> 0.8		< 0.7	14

in	out
7730	5630
6060	6530
5530	5090
33300	18440
19330	20800
21010	19370
118220	6000

Figure 1. *A gauche : la colonne |sommets| présente les sommets observés en fonction de la contrainte sur le degré entrant in et sur l'indice de chevauchement $overlap$; la colonne $\%$ correspond au pourcentage de sommets ayant un indice de chevauchement supérieur à 0.8 pour une classe de $ratio$ donnée. A droite sont illustrés les degrés entrants et sortants moyens sous les contraintes in, $overlap$ et $ratio$ à gauche.*

3.3. *Interprétation des résultats*

Comme indiqué précédemment, nous considérons uniquement les utilisateurs ayant un indice de chevauchement proche de 1, ces derniers étant de potentiels capitalistes sociaux. Lorsque le degré entrant est supérieur à 2000, nous observons bien les deux comportements **IFYFM** (61% des utilisateurs ont un ratio supérieur à 1) et **FMIFY** (28% des utilisateurs ont un ratio compris entre 0.7 et 1 ont un grand indice de chevauchement). De manière similaire, ces deux groupes se distinguent lorsque nous considérons des utilisateurs ayant un degré entrant supérieur à 10000. Sur la dernière ligne de la Figure 1, nous voyons également se dégager les capitalistes sociaux dits *passifs* : 14% des utilisateurs ont un ratio inférieur à 0.7, pour un indice de chevauchement supérieur à 0.8. Nous remarquons qu'en moyenne, ces utilisateurs ont un degré entrant considérablement supérieur à leur degré sortant - près de 20 fois plus grand.

4. Conclusion et perspectives

En utilisant une base de données adaptée, nous avons pu stocker et analyser le *graphe des spammers* de **Twitter**, contenant plus de **15 millions de noeuds** et plus d'**1 milliards d'arcs**. Grâce à différentes mesures de similarité, l'existence de plusieurs groupes de *capitalistes sociaux* dans ce graphe a été mise en avant, approfondissant ainsi des idées développées par Ghosh et al. [GHO 12]. Une suite logique de nos travaux est bien évidemment de calculer ces indices pour le graphe complet de Twitter. De plus, un travail de validation de nos résultats est en cours et semble prometteur. Celui s'effectue à partir d'une liste d'utilisateur identifiés comme capitalistes sociaux et fournie par Ghosh et al. [GHO 12]. Enfin, nous souhaitons évaluer l'impact des capitalistes sociaux sur les communautés au sein du graphe de Twitter. En effet, ces utilisateurs suivent et se font suivre dans le seul but de gagner de l'influence et obtiennent ainsi un degré élevé. Nous conjecturons donc qu'ils peuvent jouer le rôle de passerelles entre différentes communautés et ainsi les fausser.

Remerciements

Les auteurs remercient Mathieu Chapelle pour des discussions intéressantes, et Sylvain Jubertie pour nous avoir permis d'utiliser la machine SPEED.

5. Bibliographie

[CHA 10] CHA M., HADDADI H., BENEVENUTO F., GUMMADI K. P., « Measuring User Influence in Twitter : The Million Follower Fallacy », *ICWSM*, May 2010.

[GHO 12] GHOSH S., VISWANATH B., KOOTI F., SHARMA N. K., GAUTAM K., BENEVENUTO F., GANGULY N., GUMMADI K., « Understanding and Combating Link Farming in the Twitter Social Network », *WWW*, April 2012.

[MAR 12] MARTÍNEZ-BAZAN N., MUNTÉS-MULERO V., GÓMEZ-VILLAMOR S., ÁGUILA-LORENTE M., DOMINGUEZ-SAL D., LARRIBA-PEY J.-L., « Efficient Graph Management Based On Bitmap Indices », *IDEAS*, August 2012, to appear.