# Projet Informatique
# Détecter les capitalistes sociaux sur Twitter

**Ioan** Todinca**, Anthony** Perez

## 1 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.

**Représenter les réseaux sociaux.** 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 la suite de cet énoncé, nous dénoterons un graphe (non-orienté) par $G = (V, E)$, où :

- $V$ représente l'ensemble des *sommets* du graphe,
- $E$ l'ensemble des *arêtes* du graphe, reliant deux sommets.

Il est par exemple possible de représenter **Facebook** en considérant les utilisateurs comme sommets du graphe, et en mettant une arête entre deux sommets si les utilisateurs correspondant sont *amis* sur le réseau.

**Graphes orientés.** Dans d'autres réseaux sociaux, il est nécessaire de donner une *orientation* aux arêtes afin de représenter la relation unissant les utilisateurs (on parle alors d'*arc*). C'est par exemple le cas sur **Twitter**, où l'arc $uv$ modélise le fait que l'utilisateur $u$ *suit* l'utilisateur $v$, sans que l'inverse ne soit (nécessairement) vrai. Dans ce cas, on parle alors de graphe *orienté* $D = (V, A)$, $A$ représentant les arcs du graphe.

Figure 1 – Exemples de graphes (non-)orientés.

**Capitalistes sociaux.** Nous allons maintenant nous intéresser à et au comportement de certains utilisateurs de **Twitter** appelés *capitalistes sociaux*. L'objectif de ces derniers est simple : avoir un maximum d'utilisateurs les suivant (des *followers*) dans le but d'avoir le plus d'influence possible. Pour ce faire, ils utilisent deux méthodes relativement simples :

- **FMIFY** (Follow Me and I Follow You) : l'utilisateur assure à ses followers qu'il les suivra en retour ;
- **IFYFM** (I Follow You, Follow Me) : l'utilisateur suit d'autres utilisateurs (des *followees*) en espérant que ceux-ci le suivent en retour.

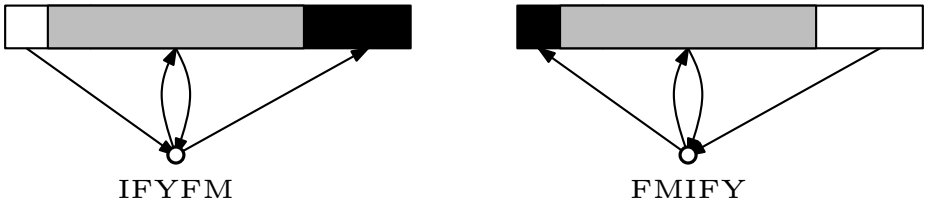


Figure 2 – Les ensembles noirs représentent les followees, les blancs les followers et les gris leur intersection.

Sur la Figure 2, le sommet blanc représente un capitaliste social. Dans le cas **IFYFM**, l'utilisateur attend d'être suivi en retour par ses followees, qui sont donc plus nombreux que ses followers. A l'inverse, dans le cas **FMIFY**,

l'utilisateur s'engage à suivre ses followers en retour, qui sont donc plus nombreux que ses followees.

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. Remarquons de plus que ces utilisateurs doivent vérifier la propriété suivante : ils ont *nécessairement* une majorité de leurs followers contenus dans leurs followees pour le principe **IFYFM**, et réciproquement pour le principe **FMIFY**.

# 2 Problématique

Ces utilisateurs ont un comportement néfaste pour un réseau social : en effet, en suivant n'importe quel utilisateur sans se soucier du contenu de ses tweets, ils permettent à des utilisateurs malsains (par exemple des *spammers*) de gagner de la visibilité, et faussent ainsi plusieurs mesures sur le réseau (à titre d'exemple, leurs tweets se retrouveront bien classés dans les moteurs de recherche, sans raison liée à leur contenu). Il est donc important d'arriver à les détecter efficacement.

Pour ce faire, vous pourrez utiliser le graphe proposé, contenant une partie du réseau social **Twitter** en **2010**. Ce dernier vous est donné sous le format suivant :

```
source        dest
source        dest
...
source        dest
```

Dans ce format, *source* et *dest* représentent des utilisateurs de **Twitter** via un *identifiant* (entier). Les valeurs de ces identifiants ne sont pas nécessairement contigües (*i.e* elles ne vont pas de $1$ au nombre d'utilisateurs).

*Remarque.* Le graphe mis à votre disposition est relativement volumineux. Il sera donc important de prendre cela en compte lors de votre implémentation.

**Détection.** Proposer et implémenter un algorithme *efficace* permettant de détecter les utilisateurs susceptibles d'appliquer les principes précédemment décrits. Cela peut être réalisé en inspectant les *voisinages* $N^+(v)$ et $N^-(v)$ de chaque sommet, correspondant respectivement à l'ensemble des followees et des followers de $v$. De plus, une seconde mesure appliquée sur ces mêmes ensembles peut permettre de *classifier* les capitalistes sociaux détectés en fonction des principes **IFYFM** et **FMIFY**.

**Vérification** Le graphe précédent datant de $2010$, il reste maintenant à s'assurer que les capitalistes sociaux détectés n'ont pas :

- été suspendus depuis,
- cessé d'appliquer les principes précédemment cités,

Pour ce faire, il est donc nécessaire de se récupérer les données nécessaires directement sur **Twitter**, et de vérifier les informations utiles en utilisant leur **API**. Il est également important de bien prendre en compte les limites d'utilisation imposées par **Twitter**.