Noname manuscript No.
(will be inserted by the editor)

Social capitalists on Twitter : detection, evolution and behavioral analysis

Nicolas Dugué · Anthony Perez

the date of receipt and acceptance should be inserted later

Abstract In this paper we focus on the detection and behavior of social capitalists, a special kind of users in Twitter. Roughly speaking, social capitalists follow users regardless of their contents, just hoping to increase their number of followers. They have first been introduced by Ghosh et al. [13] in 2012. In this work, we provide a method to detect these users efficiently. Our algorithms do not rely on the tweets posted by the users, just on the topology of the Twitter graph. Then, we show that these users form a highly connected group in the network by studying their neighborhoods and their local clustering coefficient [30]. We next study the evolution of such users between 2009 and 2013. Finally, we provide a behavorial analysis based on social capitalists that tweet on a special hashtag. Our work emphasizes that such users, who act like automatic accounts, are in fact for most of them real users.

Keywords Twitter, Social network, Social capitalism, influence

Nicolas Dugué, Anthony Perez
Université d'Orléans, LIFO EA 4022, F-45067 Orléans, France
E-mail: {anthony.perez}{nicolas.dugue}@univ-orleans.fr

1 Introduction

Context. In the last decade, large and complex data have been produced through the study of Internet [9], business intelligence [27] or bioinformatics [22]. This phenomenon, known as *Big Data*, raises a lot of research interests. In particular, being able to store, share and analyse such data efficiently constitutes a major research area [23]. In several cases, these data can be represented using graph theory. This is in particular well-suited to study social networks, where connections between users can be easily represented using graphs. Due to a huge increase in the number of users of these social networks, the graphs obtained are very large. In this paper, we consider the *relation-graphs between Twitter users* computed in 2009 [9,16] and we focus on the behavior of so-called *social capitalists*.

Twitter. Twitter is mostly used to share, seek and debate about information, or to let know the world about daily events [15]. This micro-blogging service allows users to post messages of 140 characters at most called *tweets*. Twitter includes social features and is thus considered as a social network. Indeed, to see the messages of other users, a Twitter user has to *follow* these users. And reciprocally, accounts that follow an user are able to see the messages posted by this user. Users following an account are called its *followers*. Users followed by an account are called its *friends*. Furthermore, users can *retweet* [25] messages of other users that they find to be interesting. This action allows users to spread a tweet to their followers. Besides, users can *mention* other users to draw their attention by adding *@UserName* in their message. Finally, by adding *hashtags* to their tweets, i.e. keywords preceded by a *#*, users make their tweets visible on the Twitter page dedicated to these keywords.
The online service is more and more used. Indeed, the number of user exceeds 500 millions [1] and every two and a half days, a billion tweets are sent [21]. Since Twitter became a powerful tool to spread information, the service is now used as a marketing tool by politicians, celebrities or organisations [8]. Malicious users appear too such as spammers [29] who see Twitter like a new gold mine to spread their messages. All of these users are trying to be as influent [5] and visible as possible on the network. In this paper, we focus on social capitalists, users who are trying to reach this aim by any means.

Social capitalists. These users share a common goal, that is to acquire a maximum number of *followers* to gain visibility. Indeed, the larger the number of followers of an user is, the greatest its influence in the network may be [5]. Besides this obvious interest, accumulating followers is also useful since it has a direct incidence on ranking tweets on search engines [13]. Such a behavior has first been described by Ghosh et al. [13], in a paper considering farm-linking, and especially **spammers**. The authors observed that users that respond the most to request of spammers are mainly *real users*, and they characterize this behavior by calling it *social capitalism*. These users are not healthy for a social network: since they follow users regardless of their contents, just hoping to be followed back, they may help users such as spammers to gain followers. This specific attitude provides visibility to their tweets as well, without any content-related reason. Furthermore, even if most of these users seems to be real users [13], bots and spammers may use the same techniques to gain visibility. Hence, detecting such users is of important interest since it can lead to a better understanding on the influence's *legitimacy* of a Twitter user. It is also important to analyse their behavior on the social network, and more precisely to determine whether their strategy is successful. In this case, it may have an important impact on the social network. Notice that spammers and spam campaigns were previously studied in [12,19,29] but, to the best of our knowledge, the way to detect social capitalists and how they behave remains unexplored except for the introducing paper of Ghosh et al [13].

Our contribution. We first use two similarity measures, namely the *overlap* [24] and *ratio* indices, to detect social capitalists on an anonymized Twitter graph collected in 2009 by Cha et al. [9]. To that aim, we define a threshold on the first similarity measure based on a list of certified 100.000 social capitalists detected by Ghosh et al. [13] on this graph using an ad-hoc method (Section 2). Then, we use the second measure to classify efficiently such users w.r.t. the social capitalism technique they use. It is interesting to notice that our detection algorithm relies on the topology of the Twitter graph only, and do need neither to own tweets nor mentions. After having detected the social capitalists using our measures, we show that they share some features. In a first place, we compute the local coefficient of clustering [30] of the detected social capitalists (Section 3). We observe that the coefficient of social capitalists is much higher than the coefficient of other users. Furthermore, we observe that most users following social capitalists are also social capitalists. These two observations let us know that social capitalists are a highly connected subgroup in the network. Next, we give a picture of the evolution of social capitalists between 2009 and 2013 (Section 4). To that aim, we use a public Twitter dataset crawled in 2009 by Kwak et al. [16] that contains more than 40 millions vertices and about 1.5 billions arcs. To begin, we compute the overlap and ratio indices of all these users, and then detect roughly 145.000 potential social capitalists w.r.t. the threshold previously defined. Then, with a taylor-made program based on the Twitter API, we crawled the current neighbors of 75% of these users. We compute the overlap and ratio indices on our crawl, and compare the results with those obtained with the dataset provided by Kwak et al. To conclude our analysis, we provide a behavorial analysis of social capitalists by focusing on specific hashtags dedicated to social capitalism (Section 5). To that aim, we gathered a dataset of roughly 725000 tweets posted using these hashtags and provide a qualitative study for it. By looking at the sources of these tweets, we observe that most users tweeting on these hashtags are real and post their messages manually. This confirms a behavior first observed by Ghosh et al. [13] in 2010, who emphasized that accounts that respond the most to sollicitations of spammers are indeed *real users*, and act in fact as social capitalists. Then, we created an automatic account with no profile and no other aim than getting followers. This account tweets and retweets each hour, exclusively on hashtags dedicated to social capitalism that we detected in the dataset. We observe that such a strategy allows to gain followers and thus visibility in a small amount of time.

2 Social capitalists

2.1 Definition and examples

Similarly to Internet behaviors, where websites administrators perform *links exchange* in order to increase their visibility, some social network users seek to obtain a maximum number of virtual relationships. Twitter is particularly well-suited to observe and study these kinds of behavior. Indeed, microblogging networks are focused on sharing information, not on friendship links. Therefore, it is possible to become visible on this kind of networks by accumulating followers and thus, to spread information efficiently. To achieve this goal, such users (called *social capitalists* in [13]) exploit two relatively straightforward techniques, based on the reciprocation of the *follow* link:

- **FMIFY** (Follow Me and I Follow You): the user ensures its followers that it will follow them back;
- **IFYFM** (I Follow You, Follow Me): at the contrary, these users follow other users hoping to be followed back.

Social capitalists have first been lightened by Ghosh et al. [13] during a study focused on *spammers*. Spammers are engaged in a so-called *farm linking* principle, where the aim is to be followed by a maximum number of users to spread spam links. In [13], the authors observed that users that respond the most to requests of spammers are in fact real users (*i.e.* not spammers nor false accounts). If the fact that such users are not false accounts can be surprising at first sight, it can actually be easily explained: as mentioned previously, social capitalists use the **FMIFY/IFYFM** principles, and follow back users that follow them regardless of the content of their tweets.

To give a better picture on social capitalists, we provide in Table 1 a short list of well-known social capitalists such as **Barack Obama**, **Britney Spears** or **JetBlue Airways** [13]. Other users are added to this list according to their screen name (used to identify the account when preceded by ”@”) or name explicitly associated to social capitalisms such as *TFBJP* or *iFollowBack*.

screen name	**name**	**followers**	**friends**
IFOLLOWBACKJP	TFBJP	$1.2 \cdot 10^5$	$1.1 \cdot 10^5$
itsrealchris	iFollowBack	$1.7 \cdot 10^5$	$1.6 \cdot 10^5$
AllFollowMax	TFBJP	$4.2 \cdot 10^4$	$4.3 \cdot 10^4$
BarackObama	Barack Obama	$2.5 \cdot 10^7$	$6.7 \cdot 10^5$
britneyspears	Britney Spears	$2.2 \cdot 10^7$	$4.1 \cdot 10^5$
JetBlue	JetBlue Airways	$1.7 \cdot 10^6$	$1.1 \cdot 10^5$
Starbucks	Starbucks Coffee	$3.2 \cdot 10^6$	$7.9 \cdot 10^4$

Table 1 Well-known or obvious social capitalists. We want to mention that TFBJP means *Team Follow Back Japan*. Followers and friends numbers are rounded.

2.2 Similarity measures

In order to detect social capitalists, we make use of two similarity measures, namely *overlap index* (introduced in [24]) and *ratio*. As we shall see, the first one will enable us to detect potential social capitalists, while the latter one will allow us to determine if a given user uses the **FMIFY** or **IFYFM** principle.

Definition 1 (Overlap) Given two sets A and B, the *overlap index* of A and B (which is between 0 and 1) is given by:

$$O(A,B) = \frac{|A \cap B|}{\min\{|A|,|B|\}}$$

For every vertex v, we first apply the overlap index on the sets $A = N^+(v)$ and $B = N^-(v)$ with $N^+(v)$ (resp. $N^-(v)$) the set of out- (resp. in-) neighbors of v. Because we are considering the Twitter network, A (resp. B) is the set of *friends* (resp. *followers*) of the user v. We claim that this allows us to detect users that are likely to be social capitalists. Indeed, due to the applied principles, the relation between their in- and out-neighbors must be strong, and thus most of the users they follow should follow them back. In particular, this means that their overlap index must be close to 1. We next use Definition 2 to classify more precisely such users.

Definition 2 (Ratio) Let v be any vertex of the Twitter graph. The *ratio* of v is given by:

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

Intuitively, social capitalists following the **IFYFM** principle should have a ratio greater than 1 (*i.e.* more friends than followers), while the users following the **FMIFY** should have a ratio less than 1. Recall that, in both cases, the expected ratio should be close to 1. However, as we shall see in Section 3, another behavior arises that we call *passive*. Unlike other social capitalists (called *active*), these users find their in-degree large enough, and thus estimate having enough influence. Once they have reached this point, they stop using the aforementioned principles but still get more and more followers. Consequently, their overlap index should still be close to 1 whereas their ratio should be a lot smaller than 1.

2.3 A threshold on the overlap index

As stated in Section 2.2, users with an overlap index close to 1 are likely to be social capitalists. We now define what *close to* 1 means, and provide a threshold on the overlap index that should statistically allow us to detect efficiently social capitalists.

To do so, we use a list of 100000 users of the relation-graph provided by Cha et al. [9] considered to be social capitalists with a very high probability. This list is not exhaustive and the ad-hoc method used to detect these users requires to detect first spammers, which is difficult and long. Indeed, such users have been detected by Ghosh et al. [13] in the following way: while studying the behavior of spammers in Twitter, they noticed that users that respond the most to sollicitations of spammers are actually *real users*. Hence, they ranked such users according to the number of spammers they follow, and then provided a list corresponding to the 100.000 that follow the most spammers. We thus compute their overlap index in the relation-graph as well as the distribution of the values of the overlap index among them (Figure 1).

We observe that a few users of this list have an overlap less than 0.6. We may wonder whether these users really are social capitalists. Actually, the only conclusion we can draw is that they do not use the principles described before. Above 0.6, there is a dramatic increase in the number of users for each interval. Notice that 80% of these users have an overlap greater than 0.74. Since we want to preserve users who are more likely social capitalists and to avoid too many false positives, we must use a high threshold. Due to these observations, we choose 0.74.

This threshold is very close to the one chosen in a previous work [11], which was 0.72. The difference comes from the fact that the results in [11] were obtained on a so-called *spammer-graph*, which contained roughly half the graph we consider in this work. The computed threshold was thus an underestimate, but we would like to mention that all social

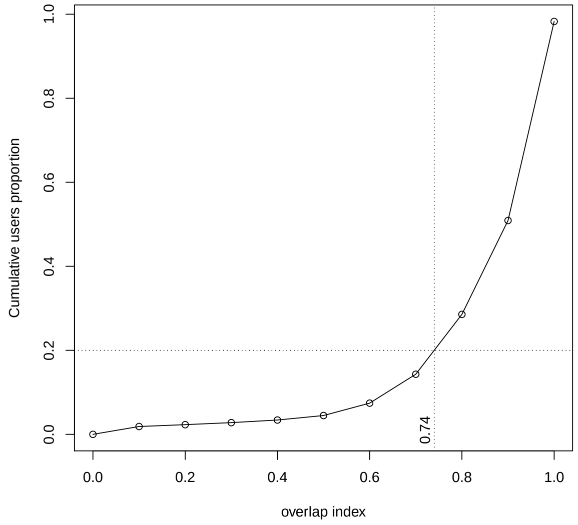


Fig. 1 Cumulative distribution of the overlap index of the 100000 potential social capitalists from the list [13].

screen name	name	overlap	ratio
IFOLLOWBACKJP	TFBJP	0.97	0.92
itsrealchris	iFollowBack	0.81	0.94
AllFollowMax	TFBJP	0.99	1.04
BarackObama	Barack Obama	0.77	0.03
britneyspears	Britney Spears	0.82	0.02
JetBlue	JetBlue Airways	0.74	0.06
Starbucks	Starbucks Coffee	0.77	0.02

Table 2 Overlap and ratio of the social capitalists of Table 1 obtained using the Twitter API. Values are rounded.

capitalists detected in the spammer-graph by using this previous threshold are still considered as social capitalists in our new framework.

In a previous subsection, we introduced a short list of famous social capitalists with the Table 1. Using our similarity measures and the threshold, we can detect these users. Indeed, their overlap index is greater than 0.74 as shown in Table 2. The first two users are considered as **FMIFY** users because their ratio is less than 1. The third one is classified as a **IFYFM** with a ratio greater than 1. We observe that the ratio of the four last ones is really close to 0, implying that they are passive social capitalists, which makes sense since they correspond to famous Twitter accounts.

3 Detection and characterization

3.1 Detection

In Section 2.3, we determined a threshold on the overlap index which is 0.74. With this threshold, we are now able to detect social capitalists on the whole Twitter graph provided by [9]. This graph contains roughly 55 millions vertices which represent Twitter users. The almost 2 billions arcs represent the follow links between users. We processed the overlap index and the ratio on every user of the dataset. Furthermore, we introduce two other constraints. The first one is on the number of followers. We are interested in successful social capitalists who effectively gained followers applying the priciples described in Section 2. We thus consider users with at least 500 followers. Besides, to avoid detecting users with a high overlap index but with a really small number of friends, we also force the number of friends to be greater than 500. Indeed, an user with a few friends and a lot of followers may have an overlap close to 1 but this case would not fit with our purpose. To summarize, the users we are considering here as social capitalists are users with an overlap index greater than 0.74 and a number of followers and friends both greater than 500.

in	**out**	**vertices**	**ratio**	%
> 500	> 500	161424		
			> 1	68
			$[0.7; 1]$	25
> 2000	> 500	47221		
			> 1	61
			$[0.7; 1]$	31
> 10000	> 500	5743		
			> 1	66
			$[0.7; 1]$	25
			< 0.7	9

Table 3 Detecting social capitalists on the entire graph. Overlap index is always greater than 0.74.

We detect a bit more than 160.000 social capitalists on the graph, which is a little more than 55% of the number of users with more than 500 followers and friends. To give a better insight about these users, we then classify users using the in-degree, namely the number of followers and the ratio. As we can see on Table 3, most users with a number of followers greater than 500 are **FMIFY** users with a ratio greater than 1. This remains true when considering users with a number of followers greater than 2000 and 10000. In the last case, we observe *passive* social capitalists, with a ratio smaller than 0.7. They represent 9% of the users with an in-degree that high. Another interesting observation is that the average ratio of this category of users is roughly 0.25. This means that in average, these users have 4 times more followers than friends.

3.2 Structural properties of the neighborhoods of social capitalists

We now want to characterize the structure of social capitalists in this graph. We know that social capitalists use the two principles **FMIFY** and **IFYFM** to increase their number of followers. Intuitively, we claim that these principles are especially efficient when the users targetted are social capitalists too. To confirm this hypothesis, we look at the neighbors of social capitalists and we observe that a large majority of them are detected as social capitalists by our method. As we can see on Figure 2, a bit less than 70% of social capitalists are followed by at least 50% of social capitalists. This shows that for a large number of social capitalists, the visibility obtained by increasing their number of followers is due to the presence of other users applying this technique. These users thus acquire a completely artificial visibility on the graph.

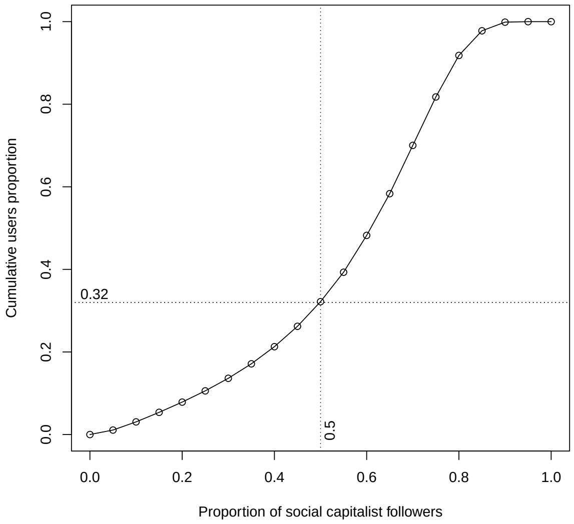


Fig. 2 Proportion of social capitalists among the followers of social capitalists.

Furthermore, we show that this has an impact on the local clustering coefficient [30] of social capitalists. We do not consider the orientation of the graph to process it. Local clustering coefficient measures the density of edges between the neighbors of a node and is thus defined for any vertex v as the fraction of its neighbors that are connected.

Definition 3 (Clustering coefficient) Let $G = (V,E)$ be any undirected graph, and $v \in V$ be any vertex. The clustering coefficient cc_v of a vertex v is defined as :

$$cc_v = 2n_{edges}/(d_v(d_v - 1))$$

where n_{edges} is the number of edges between the neighbors of v, and d_v the number of neighbors of v.

We now compare the local clustering coefficient of users that we detect as social capitalists and other users. We observe that 289636 users have more than 500 followers and friends on our graph, and 161424 of them are detected as social capitalists by using our method. Thus, 128212 of them are not considered as social capitalists. By processing the clustering coefficient on this two groups of users, we can see (Figure 3) that the clustering coefficient of social capitalists seem to be higher. In average, it is roughly twice as high as (0.084) for the users that are not considered as social capitalists (0.044).

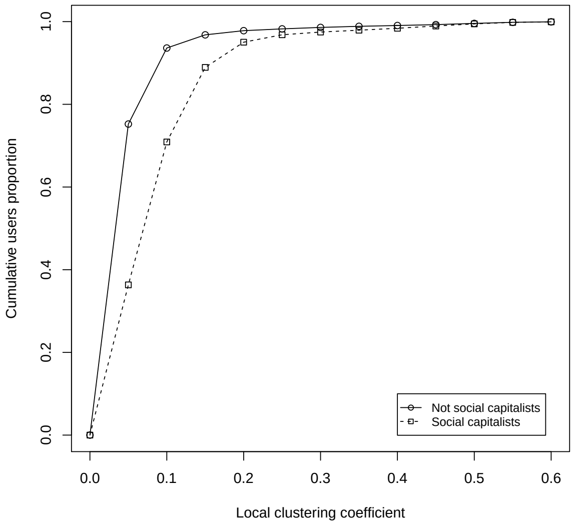


Fig. 3 Comparison of the local clustering coefficient of social capitalists and normal users .

These observations allow us to state that social capitalists constitute a densely connected subgroup inside the Twitter graph. Most of their neighbors are social capitalists, who are also more connected in average, as the higher clustering coefficient of these users showed us.

4 Evolution of social capitalists between 2009 and 2013

We now describe the evolution of social capitalists between July 2009 and July 2013. To realize such a study, we make use of a public Twitter dataset crawled in 2009 by Kwak et al. [16]. We also had to program a new crawl script using the Twitter API V1.1 [2]. Indeed, Twitter deprecated the API V1 which cannot be used anymore. We will first give some insights on both datasets, and then present the results we obtained.

Dataset. From July 6th to July 31st, Kwak et al. [16] *crawled* Twitter using the following method: they started crawling from **Perez Hilton**'s account (who had about one million followers at this time) and then crawled breadth-first along the direction of followers and friends. Doing so, they computed a directed graph containing more than 1.4 billions relations between 41 millions users of Twitter. To begin our analysis, we first compute the overlap for every user of this graph (*Twitter*'09 in what follows). Out of the 41 millions users, about 7.500.000 have an overlap greater than or equal to 0.74. Recall that such users are most likely to be social capitalists (Section 3). To obtain more relevant results, we only consider users that have both a number of followers and friends greater than 500. Recall that such a restriction prevents the detection of users with an overlap close to 1 but with only a few friends or followers (in which case the overlap index is not meaningful). Doing so, we finally obtain a list of approximatively 145.000 users. In the following, we thus focus on such users.

A new crawl. We use the fact that the data contained in *Twitter*'09 is non-anonymized. Thus, by using Twitter's API, one can retrieve the screen name of any user of the graph from its identity number[1]. In order to analyze the evolution of social capitalists between July 2009 and July 2013, we thus designed a program that crawls Twitter using the new API. Using a few Twitter accounts authenticated to the API we created, we managed to crawl the Twitter graph at the rate of approximatively 200 requests per hour, which allows to obtain at most 1.000.000 friends and followers per hour. More precisely, we launched our crawl algorithm on the list of aforementioned social capitalists in *Twitter*'09, and crawled their followers and friends. Notice however that this program does not do the same thing than the one designed by Kwak et al [16], since it starts from a given list of users instead of doing a breadth-first traversal. In what follows, we will refer to this graph as *Twitter*'13. So far, we got the information needed for more than 110.000 users, which gives us a reliable sample of more than 75% of social capitalists to analyse.

General observations. We begin by comparing the evolution of the number of followers and friends and both the overlap and ratio indices for those users (Figures 4 and 5).

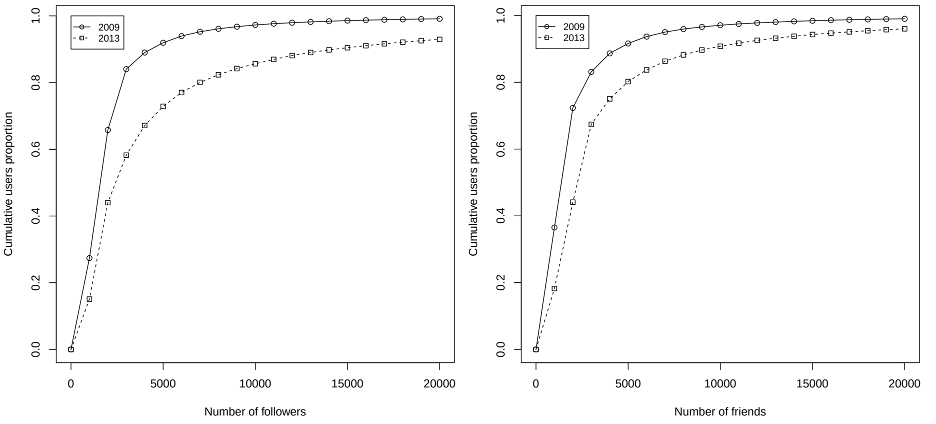


Fig. 4 Difference between the cumulative number of followers and friends distributions in *Twitter*'09 and *Twitter*'13.

As one can observe, most users have a larger number of followers and friends in *Twitter*'13 than in *Twitter*'09. This fact is not very surprising, since four years have passed between the two datasets. Furthermore, we notice that the number of followers increased more than the number of friends. The most interesting part lies in the ratio and overlap indices. Indeed, regarding the former, one can see that there are a lot of users in *Twitter*'13 that have a

[1] We would like to mention that Kwak et al. [16] also provide a correspondance between identity numbers and screen names.

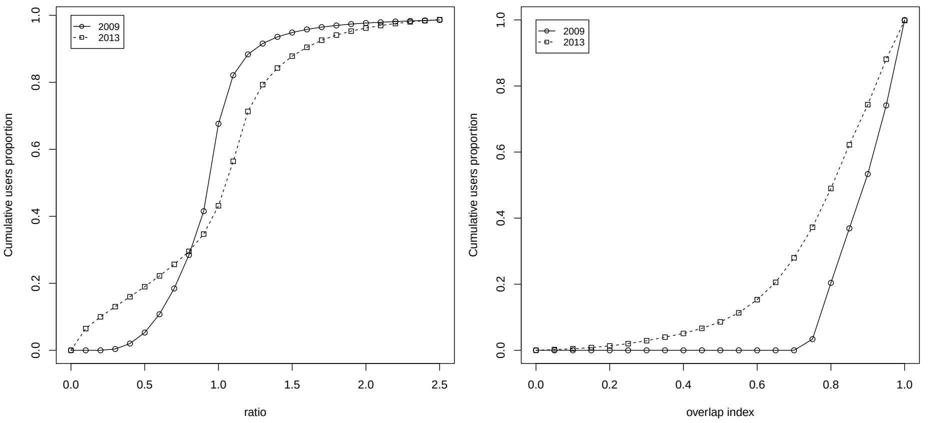


Fig. 5 Difference between the cumulative overlap and ratio distributions in *Twitter'*09 and *Twitter'*13.

ratio really close to 0, while such users almost did not exist in *Twitter'*09. On the other hand, one can observe that about 75% of the 2500 social capitalists of *Twitter'*09 that had a ratio greater than 2 now have a ratio close to 0 in *Twitter'*13. Such users thus represent social capitalists that applied the techniques successfully. It seems important to notice that, to gain their visibility, they followed users massively in 2009, and waited to be followed back in return. Such an aggressive strategy is not available anymore, since Twitter imposes a limitation on the ratio for any user [3]. This raises the question of the relevance of the influence of an user in Twitter. Regarding the overlap index, at least 50% of the users still have an overlap index greater than 0.74, and only a few users (about 10%) have an overlap index smaller than 0.5. We will present in what follows the reason for this.

Twitter'13 *users with overlap* ≥ 0.74 We now present more precisely the evolution of the ratio for the social capitalists we detected in 2009 that preserved a large overlap index in 2013. The results are presented Table 4, and show the distribution of the ratio of an user in 2013 (represented by the columns) that had a given ratio in 2009 (represented by the lines).

	number	$\mathbf{o}^{2013} \geqslant 0.74$	$\mathbf{r}^{2013} > 1$	$\mathbf{r}^{2013} \in [0.7;1]$	$\mathbf{r}^{2013} < 0.7$
$\mathbf{r}^{2009} > 1$	35058	19892	40%	25%	35%
$\mathbf{r}^{2009} \in [0.7;1]$	53986	37789	80%	14%	6%
$\mathbf{r}^{2009} < 0.7$	20026	10435	89%	7%	4%

Table 4 Statistics of the social capitalists that preserved a large overlap index. The notations r^i and o^i respectively denote the ratio and overlap indices for $i \in \{2009, 2013\}$. Percentages are related to the figures in the second column.

As one can see, these results enlight the observations that were previously done. In particular, it shows that most users of *Twitter'*09 who had a ratio greater than 1 and a large overlap maintain a ratio close to 1, meaning that they keep applying the **IFYFM** principle.

Another observation is that about 68% of these vertices have a number of followers smaller in *Twitter'*13 than in *Twitter'*09. This may indicate that they have been suspended by Twitter at some point due to their behavior or that their followers have been suspended. Twitter limitations on the ratio index probably also stopped their progression at some point. Finally, they may have been unfollowed by some users who prefered to keep their ratio low. However, for 35% of the vertices which had a ratio greater than 1 and a large overlap, the **IFYFM** method was succesfull since they now have a ratio less than 0.7 and are passive social capitalists.

Twitter'13 *users with overlap* < 0.74 We now focus on users that had a large overlap index in 2009 but that now have an overlap index smaller than the defined threshold. There are about 38000 such users, which represent roughly speaking 35% of the sample we are studying. As previously, we analyze the evolution of the ratio of such users (Table 5).

	number	$\mathbf{o}^{2013}$<0.74	$\mathbf{r}^{2013} > 1$	$\mathbf{r}^{2013} \in [0.7;1]$	$\mathbf{r}^{2013} < 0.7$
$\mathbf{r}^{2009} > 1$	35058	14103	16%	25%	59%
$\mathbf{r}^{2009} \in [0.7;1]$	53986	15377	36%	23%	40%
$\mathbf{r}^{2009} < 0.7$	20026	8894	61%	14%	25%

Table 5 Statistics of the accounts that used social capitalism techniques successsfully. The notations r^i and o^i respectively denote the ratio and overlap indices for $i \in \{2009, 2013\}$. Percentages are related to the figures in the second column.

A lot of users that had a ratio greater than 1 in *Twitter'*09 now have a ratio smaller than 0.7, which indicates that they applied the techniques successfully. Indeed, by applying the **IFYFM** principle, they gained enough followers and then may have unfollowed their friends. As one can see on Table 6, some of these users maintain an overlap quite close to the threshold. Actually, the average overlap index of these users is 0.56. This means that they keep most of the reciprocate links they had created when they were applying the **IFYFM** principle. Notice that among the 38000 social capitalists that have a small overlap index in *Twitter'*13, only 35% have an overlap index smaller than 0.56. A similar observation is made for social capitalists that had a ratio between 0.7 and 1 in *Twitter'*09.

Username	**out-**	**in-**	**ratio**	**ovp**	**out-**	**in-**	**ratio**	**ovp**
@Coldplay	1517067	2633	576.174	0.886	2391	10570550	0.000	0.701
@UberSoc	4090	1213	3.372	0.839	2418	10495065	0.000	0.570
@SHAQ	1843561	563	3274.531	0.838	1110	7225068	0.000	0.686
@ESPN	138982	109377	1.271	0.843	360	6441694	0.000	0.606
@funnyordie	593981	1691	351.260	0.782	4248	6200508	0.001	0.546
@noaheverett	3635	588	6.182	0.767	1227	4497421	0.000	0.737
@MLB	133953	8509	15.743	0.781	2198	2964578	0.001	0.602
@TFL	71027	1506	47.163	0.764	1730	2882159	0.001	0.602
@addthis	14093	13363	1.055	0.927	22762	2750610	0.008	0.736
@TechCruch	1041057	691	1506.595	0.839	863	2728905	0.000	0.718

Table 6 Statistics of the 10 first accounts that have a small overlap index ranked by their number of followers in 2013. The numbers for *Twitter'*09 are on the left, the ones for *Twitter'*13 on the right.

We now consider users with a relatively small overlap index (less than 0.25) in *Twitter'*13. There are about 2500 such users. We observe that half of them had a ratio close to 1 in *Twitter'*09, but now have a really small ratio (less than 0.23). Since their overlap index is also very small, these users correspond to social capitalists that applied the principles successfully, and then decided to unfollow massively their friends. Massively unfollowing friends is an efficient technique as we will see in Section 5.4. On the other hand, if we consider users with a ratio smaller than 0.7 in *Twitter'*09, we observe that the difference between their number of friends and followers in both graphs is relatively small. Hence, they might correspond to users that stopped applying the principles, and became *regular* users.

Focus on particular accounts. To conclude this part of our work, we would like to focus on several accounts who succeeded using social capitalism techniques. More precisely, we consider accounts who had both an overlap index greater than 0.74 and a ratio greater than 1 in *Twitter'*09, and that still have a large overlap index in *Twitter'*13 but with a significantly smaller ratio (namely lower than 0.7). An interesting observation from our results is that the first accounts who gained the more followers using this technique are respectively **Lady Gaga**, **Barack Obama**, and **Britney Spears**, who constitute well-known social capitalists [13]. We present similar significant results in Table 7, where some accounts present a dramatic improvement in their ratio (e.g. **paulpierce34**, who went from a ratio close to 1500 to a ratio close to 0). As the accounts that did not preserve a large overlap index but yet succeeded using social capitalism techniques, those particular accounts used the fact that Twitter did not impose a limitation on the ratio at this time. Once they have reached a point where they could gain followers without applying the techniques anymore, they unfollowed massively their friends.

Username	**out-**	**in-**	**ratio**	**ovp**	**out-**	**in-**	**ratio**	**ovp**
@ladygaga	636929	73274	8.69	0.97	136386	37485540	0.00	0.82
@BarackObama	1882889	770155	2.44	0.91	680428	30836226	0.02	0.77
@BritneySpears	2674874	406238	6.58	0.95	412703	27763836	0.01	0.81
@paulocoelho	75423	48446	1.56	0.98	98	7721670	0.00	0.86
@Anahi(Magia)	15337	1765	8.69	0.96	455	6492119	0.00	0.90
@stephenfry	693512	55044	12.60	0.90	54122	5823567	0.01	0.81
@hootsuite	80936	61828	1.31	0.94	1274698	5013919	0.25	0.75
@TheOnion	1380160	369569	3.73	0.87	8	4931732	0.00	0.87
@showdavida	598444	1005	595.47	0.90	36	4651921	0.00	0.75
@yokoono	81765	71585	1.14	0.95	1003791	4311890	0.23	0.79
@DwightHoward	727413	2564	283.70	0.82	8037	4134991	0.00	0.84
@Starbucks	271215	138045	1.96	0.96	86594	3723806	0.02	0.77
@NatGeo	14339	11755	1.22	0.99	23792	3668447	0.01	0.88
@WholeFoods	1112628	498700	2.23	0.89	562219	3400523	0.16	0.76
@jimmycarr	183925	1344	136.85	0.89	228	3302043	0.00	0.74
@wossy	386479	3985	96.98	0.96	5966	3249585	0.00	0.87
@wyclef	643237	3412	188.52	0.93	8649	3246278	0.00	0.80
@paulpierce34	815197	524	1555.72	0.95	85	2804060	0.00	0.78
@zappos	1075935	407705	2.64	0.88	380335	2759301	0.14	0.75
@Arsenal[DotCom]	6282	5538	1.13	0.98	165258	2545893	0.06	0.86

Table 7 Statistics of the accounts that used social capitalism techniques successsfully. The numbers for *Twitter'*09 are on the left, the ones for *Twitter'*13 on the right.

5 Behavioral analysis: the #TeamFollowBack hashtag

We showed previously that social capitalists use the two **FMIFY** and **IFYFM** principles to increase their number of followers and thus be as visible as possible. In this section, we give an insight about how these users may apply these methods in practice. We may assume that some of them follow the neighbors of their followers, or that other ones follow the users suggested by Twitter. But that would be underestimating the organization of such users. Indeed, part of the social capitalists are gathered in teams such as *TeamFollowBackJapan*, whose user *IFOLLOWBACKJP* (see Table 1) seems to be a member. These teams often use a dedicated hashtag to apply their methods. Users that tweet on these hashtags are part of the team or likely to be social capitalists. Actually, tweeting on these hashtags allow to find potential social capitalists to follow. Several methods may be used by these users such as:

- tweeting messages inviting other users to follow them;
- tweeting messages inviting other users interested in being followed by them to retweet them (see Figure 6);
- tweeting messages mentioning users they want to be followed by;
- follow users that are applying all these methods on this hashtag;

To study these methods, we crawled the hashtag *#TeamFollowBack* during the month of February 2013 using the streaming API provided by Twitter. We eliminated the redundant tweets and finally obtained a large tweets dataset of more than 725000 tweets.

5.1 Hashtags

To introduce the data, let us give a few global figures about the tweets gathered on the *TeamFollowBack* hashtag.

Type	**Number**
Tweets	726470
Hashtags	4227703
Distinct hashtags	25028
Average hashtag number by tweet	5.82
Distinct users	124786
Mentions	719972
Distinct user mentioned	43199

Table 8 Statistics about the tweets crawled on the hashtag *#TeamFollowBack*.

As we can see in Table 8, there are almost 6 hashtags per tweet in average. This is a high rate, even bigger in average than the one observed for the spammers tweets in [6]. Consequently, we took a deeper look to the hashtags used by the social capitalists in these tweets. To begin, they used in this dataset 25028 distinct hashtags. But only a few of them are really frequent. Indeed, by summing the occurence number of the ten most used hashtags of the dataset, we obtain more than 50% of the number of hashtags occurences in the dataset. Doing the same with the 46 most used hashtags, we obtain more than 90% of the number of hashtags occurences in the dataset. To sum this up, by looking at 0.2% of the hashtags in our dataset, we can study more than 90% of the hashtags occurences in the dataset.

Hashtag	Occurence numbers
TeamFollowBack	766421
TFBJP	339917
sougofollow	177064
500aday	172655
OPENFOLLOW	148304
FollowBack	143174
RT	125211
instantfollowback	107989
AutoFollow	105528
HitFollowsTeam	102100
90sBabyFollowTrain	100742
f4f	100043
mustfollow	100041
FollowNGain	99967
follow	92415
tfb	83820
FF	82299
1000aday	79657
teamautofollow	59670
autofollowback	53479
IFollowBack	53032
maxvip	51316
followme	49809
retweet	48656
rt2gain	48008
teamfollowwack	43856
follow2gain	41499
TFW	40637
teamhitfollow	35405
Followers	33546
hdyf	28750
thf	27861
INSTANTFOLLOW	27075
R_Family	25293
teamfollowwacky	22495
Retweets	21384
followmejp	21340
Favorites	19655
love	19162
tmw	18087
follow4follow	17346
SiguemeyTeSigo	16982
RTRTRT	16161
tfbfollowtrain	15784
OpenFollowTeam	14590
ffback	14576

Table 9 The 46 most used hashtag and their occurence numbers in the dataset.

We observe on Table 9 that most of these hashtags are explicits. The keywords *team*, *follow* (often abbreviated *f* as in *TFBJP*), *back*, *retweet* (often abbreviated *rt* as in *rt2gain*) are present in almost all hashtags. We can see some exceptions such as *love* and *Favorites*, which may be popular hashtags as the ones used by spammers [6].

5.2 Sources of the tweets from *#TeamFollowBack*

By gathering tweets from the hashtag, we were also able to get the sources used by the users to post their messages. This information is important because it indicates whether or not these users are automating the way they are tweeting on this hashtag.

Among the 606 distinct sources used to post the tweets obtained in this dataset, 15 are particularly relevant because 90% of the tweets were posted using them. As we can see on Table 10, some of these sources are websites and applications that allow to automate some functionnalities on a Twitter account, and especially tweeting. However, only 12% of the tweets gathered are posted using these sources. Thus, most tweets are posted using official Twitter applications for phones and devices or applications to manage a twitter account in a more user-friendly way. Most of these users are neither bots nor automated accounts, but real users that are applying social capitalism in a human way.

5.3 Overlap index and ratio of users

Among the 124786 distinct users who tweeted on the *#TeamFollowBack* hashtag, we crawled the followers and the friends of 12% of the initial users. With these data, we computed the

Sources	Occurence numbers
<a href="http://blackberry.com/twitter">Twitter for BlackBerry</a>	196911
<a href="http://twitter.com/">web</a>	196747
<a href="http://twitter.com/download/android">Twitter for Android</a>	66473
<a href="http://twitter.com/download/iphone">Twitter for iPhone</a>	49556
<a href="http://twitterfeed.com">twitterfeed</a>	39010
<a href="https://mobile.twitter.com">Mobile Web (M2)</a>	28482
<a href="http://botmaker.dplays.net/">BotMaker</a>	14260
<a href="http://app.bestfollowers.me">BestFollowers App.2.85</a>	11189
<a href="http://dlvr.it">dlvr.it</a>	10443
<a href="http://www.tweetcaster.com">TweetCaster for Android</a>	9777
<a href="http://twitter.com/#download/ipad">Twitter for iPad</a>	9329
<a href="http://twittbot.net/">twittbot.net</a>	8300
<a href="http://ubersocial.com">UberSocial for BlackBerry</a>	8037
<a href="http://www.writelonger.com">Write Longer</a>	5155
<a href="http://www.twisuke.com">twisuke</a>	5040

Table 10 Sources used to post 90% of the dataset tweets. Sources allowing to automate Twitter activities are indicated in bold.

overlap index of these users in order to check whether these users are social capitalists in the sense of our definition.

Type	Number
Users observed tweeting on the hashtag	124786
Sample with their followers and friends crawled	15226
Sample users with followers and friends >500	8442
Sample users with followers and friends >500 and overlap >0.74	6740

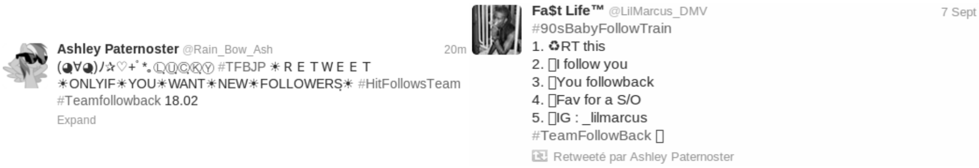
Table 11 Number of users tweeting on the *#TeamFollowBack* hashtag.

As we can see on Table 11, more than half of them have a number of followers and friends less than 500. This confirms the hypothesis that most of the users did not automate the way they are capitalizing followers. Doing it manually makes this a slow process and so, a lot of users have still a low number of followers. But 80% of the users having more than 500 followers and friends have an overlap index greater than 0.74.

5.4 Automatic social capitalism

To validate the method, we designed an automatic social capitalist account. As the bot created by [4] on *aNoobi.com*, a social network dedicated to book amateurs, we aim to show that an untrustworthy account with no profile, no content and no other aim than getting followers may succeed in obtaining followers and thus visibility on the network. Therefore, only very straightfoward functionnalities were implemented. The account is registered under the name *@Rain_bow_ash*, and can thus be checked at the following URL: *http://www.twitter.com/Rain_bow_ash*.

Each hour, the account we call *bot* posts a tweet that we choose among those of the dataset we built by crawling tweets on the *#TeamFollowBack* hashtag. Furthermore, each hour, it retweets a message originally posted on the hashtag *#90sBabyFollowTrain*, because these messages match with the messages we listed at the beginning of the Section (see Figure 6):



Ashley Paternoster @Rain_Bow_Ash 20m
(◕∀◕)/☆♡+˚*｡ⓁⓊⒸⓀⓎ #TFBJP ☀ R E T W E E T ☀ONLYIF☀YOU☀WANT☀NEW☀FOLLOWERȘ☀ #HitFollowsTeam #Teamfollowback 18.02
Expand

Fa$t Life™ @LilMarcus_DMV 7 Sept
#90sBabyFollowTrain
1. ♻RT this
2. □I follow you
3. □You followback
4. □Fav for a S/O
5. □IG : _lilmarcus
#TeamFollowBack □
Retweeté par Ashley Paternoster

Fig. 6 Examples of tweet and retweet posted by the bot using the aforementioned hashtags.

Then, by using the Twitter API, we designed several triggers to answer to interactions of other users with the bot:

- each time the bot is mentioned, it follows the account that mentioned it;
- each time it is retweeted, it follows the account that retweeted it;
- each time it is followed, it follows back the account.

We started the bot on May 30^{th}. It was stopped for a week because of utilities problems.

Type	Number
Current number of friends	7124
Current number of followers	6792
Followers number for the whole period	9319
Suspended followers	692

Table 12 Statistics about the friends and followers of the automated account doing social capitalism.

These figures were gathered on August, 23^{rd}. As we can see on Table 12, the bot is followed by 6792 users and follow 7124 users. However, our bot was followed by 9319 distinct users during the whole period of the experiment. Some of these users stopped following the bot, but a lot of users who are not followers of the bot anymore are also suspended users or deleted accounts. More surprisingly, the bot was followed 9753 times. This means that some users stopped following us and then followed us again. We observe that 324 users had this behavior, one of them even followed the bot 12 times. We assume that these users are social capitalists who are trying to cheat with the principles. They follow the bot, and then stopped following it to keep their ratio low. The first benefit of keeping a low ratio is that Twitter enforces restrictions on the ratio when an account have more than 2000 followers [3]. Thus, unfollowing users allow to follow new users and so, to increase in a quicker way the number of followers. The second benefit is for the account image. Having a bigger number of followers than friends make think the users that see the account that it is influent. We now study the efficiency of different strategies used by social capitalists.

Being retweeted and mentioned. As one can see on Tables 13 and 14, the several strategies implemented by our bot have a different level of efficiency. Some of them clearly lead to gain a higher number of followers. For instance, the bot is highly retweeted, and being retweeted by an user leads to make this user a new follower in 50% of the cases. At the contrary, the bot is more mentioned after being followed. It seems that being mentioned is not really a way to ask for a new mutual follow link since it is only the case for 23% of the mentions.

	users	**distinct users**	**follow**	**percentage**
mention	977	646	146	23
retweeet	20526	5728	2893	50

Table 13 Proportion of users that followed the bot after having retweeted or mentioned him.

Retweeting and mentioning. Retweeting and especially mentioning users is a lot less efficient. Only 11% of mentions lead to be followed by the user mentioned. Mentioning user may be less efficient because the messages we are tweeting with mentions are not up-to-date. The users we mention may thus be not interested anymore in this kind of process.

	users	**distinct users**	**follow**	**percentage**
mention	877	234	26	11
retweeet	3527	1331	470	35

Table 14 Proportion of users that followed us after being mentioned or retweeted.

Tweeting. Eventually, it seems that the simplest strategy is the most efficient. Indeed, basically tweeting on the hashtag allows our bot to gain 5784 followers, which is more than all the other strategies (3535). We can state that it is the visibility of our tweets on these hashtags who lead to be followed. Indeed, we stopped tweeting four days and gained much less followers. This confirms that the users that follow the bot without any relations with a retweet or a mention, follow him only because they saw tweets of the bot. It excludes the possibility that these users follow the bot because one of their followers follow it or any reasons close to this one.

Regarding visibility, after roughly 55 days of experiments, we can observe a dramatic increase in the number of times the bot was retweeted and the number of new followers it gained per day (Figure 7). We assume that because its number of followers increased, its tweets and retweets began to be more visible on the hashtags and thus, more users interacted with it. This hypothesis seems coherent with the fact that when the bot stopped tweeting, its number of followers barely increased.

We can conclude from these facts that tweeting on these hashtags allows to be retweeted and mentioned which leads to gain followers. But, the simple fact to be visible on these hashtags is the most efficient way to accumulate followers, because of the social capitalists using the **IFYFM** principle. We would like to mention that Twitter did not block us once while it does seem pretty obvious that the behavior of the bot is automatic.

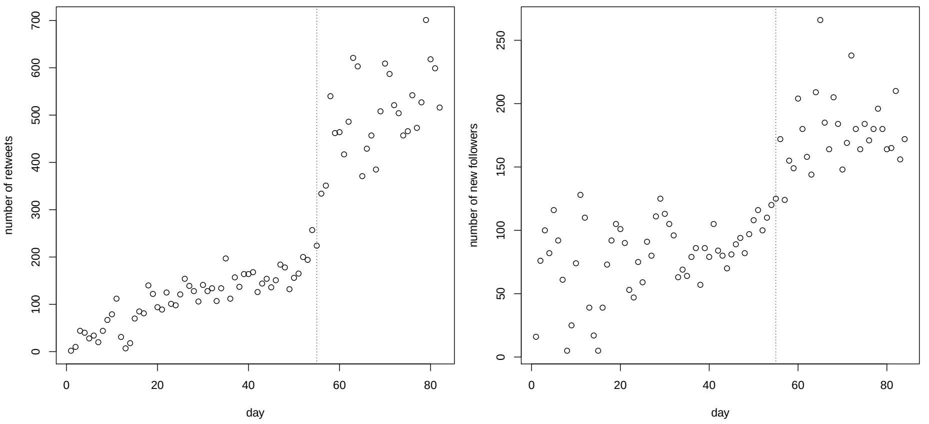


Fig. 7 Daily number of retweets (at left) and new followers (at right) during the experiment

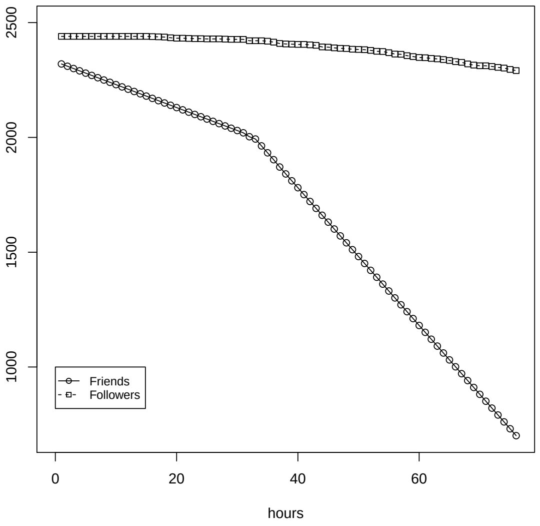


Fig. 8 Evolution of the number of friends and followers of an user unfollowing 10 friends per hour and then 30 friends per hour.

Unfollowing friends. Finally, to validate the efficiency of unfollowing friends to keep a low ratio, we designed a program which allows an user to unfollow its friends. We runned this program using one account with almost as many friends (2322) as followers (2440). This account is another bot we created to study the impact of unfollowing friends. The friends of this account were detected as social capitalists. The program unfollowed 10 friends per hour during 30 hours. Then, it unfollowed 30 friends per hour during 46 hours. As we can see on Figure 8, most of the users that are unfollowed by an user do not unfollow this user reciprocally. The account unfollowed 1680 users in a bit more than three days, and less than 150 of these users unfollowed the account. This confirms again that most of these accounts are not automated but administrated manually. Thus, it is possible to use the $iFYFM$ principle but to finally keep a low ratio.

6 Conclusion

In this paper, we focused on special users of Twitter called *social capitalists*, whose aim is to use different techniques to obtain as many followers as possible. We focus on three main techniques they use: the so-called principles **FMIFY** and **IFYFM** and the use of popular hashtags in tweets to reach as many users as possible. Such users are not healthy for social networks, since they connect to other users without any content-related reason. We first prove that these social capitalists can be efficiently detected and characterized using two similarity measures, namely overlap and ratio indices. We provide a threshold on the first measure to detect social capitalists, and then use the second one to classify them. Next, we analyze the social capitalists detected and show that they share some features. In particular, we consider their neighborhoods and process their local coefficient clustering to show that social capitalists are highly connected between them. We also consider the evolution of social capitalists between July 2009 and July 2013, using a public dataset for the first one [16] and providing a new program to crawl Twitter for the second one. We observe that most users detected as social capitalists in 2009 succeedeed in gaining followers using the aforementioned principles. More precisely, a large number of users having more friends than followers in 2009 have the opposite ratio nowadays. We also detect a large number of famous Twitter accounts that applied this principle to gain followers, and can now rely on this base to carry on growing. These accounts include the ones of **Barack Obama**, **Lady Gaga** and **Britney Spears**, who are well-known social capitalists [13]. Finally, we studied a social capitalism method based on hashtag dedicated to gain followers. We showed that most of the tweets posted on these hashtags are posted manually. Still, by using these tweets, we designed an automated account able to apply social capitalisms methods and thus to gain followers efficiently without being suspended by Twitter.

We believe that our research may lead to other interesting research questions about social networks. First of all, we think that social capitalists exist on other social networks. It would be interesting to check if this hypothesis is true and to determine why these users are present on some social networks and not on others. Then, to better undestrand the structure and the organisation of these users, we want to analyze the network communities w.r.t. these users. It seems natural to believe that they should play a special role in community detection, since they have a high local clustering coefficient and high degrees. In this sense, Dugué et al. [10] recently started a study of the *roles* of social capitalists on Twitter, using methods of clustering and measures similar to the ones provided by Guimerà et al [14]. This provides promising results and gives a better insight on the roles played by social capitalists in communities.

Furthermore, it would be interesting to refine the existing ranking method such as TwitterRank [31], or CollusionRank [13] by taking into account social capitalism through measures like the overlap index. Indeed, social capitalists gain followers artificially and so, they do not deserve the visibility that such algorithms would provide them.

Finally, another main research question would be to quantify the relationship that may exist between the number of followers, the visibility on the network and the real influence. For instance, we may wonder whether automatic accounts like the one described in the article can be used to control spam campaigns like the ones described in [26].

Acknowledgements Nicolas Dugué benefits from a PhD grant of the *Conseil Général du Loiret*. This work is partly supported by the INEX project (http://traclifo.univ-orleans.fr/INEX) funded by the *Conseil Général du Loiret*. The authors would like thank Sylvain Jubertie for its technical help, and Joris Falip who provided the source code for the bot and the crawlers during its internship.

References

1. The Telegraph : Twitter in numbers (2013). http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html
2. Twitter API V 1.1 overview (2013). https://dev.twitter.com/docs/api/1.1/overview
3. Twitter support center : Why can't I follow people? (2013). https://support.twitter.com/groups/52-connect/topics/213-following/articles/66885-why-can-t-i-follow-people
4. Aiello, L.M., Deplano, M., Schifanella, R., Ruffo, G.: People are strange when you're a stranger: Impact and influence of bots on social networks. In: ICWSM. The AAAI Press (2012)
5. Bakshy, E., Hofman, J.M., Mason, W.A., Watts, D.J.: Everyone's an influencer: quantifying influence on twitter. In: Proceedings of the fourth ACM international conference on Web search and data mining, WSDM '11, pp. 65–74 (2011)
6. Benevenuto, F., Magno, G., Rodrigues, T., Almeida, V.: Detecting Spammers on Twitter. In: Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS) (2010)
7. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. J. of Stat. Mech.: Theory and Experiment **2008**(10), P10,008 (2008)
8. Burton, S., Soboleva, A.: Interactive or reactive? : marketing with twitter **28**, 491–499 (2011)
9. Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P.: Measuring User Influence in Twitter: The Million Follower Fallacy. In: ICWSM '10: Proc. of int. AAAI Conference on Weblogs and Social (2010)
10. Dugué, N., Labatut, V., Perez, A.: Rôle communautaire des capitalistes sociaux dans twitter. In: MARAMI (2013). To appear
11. Dugué, N., Perez, A.: Detecting social capitalists on twitter using similarity measures. In: Complex Networks IV, *Studies in Computational Intelligence*, vol. 476, pp. 1–12 (2013)
12. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement, IMC '10, pp. 35–47 (2010)
13. Ghosh, S., Viswanath, B., Kooti, F., Sharma, N.K., Korlam, G., Benevenuto, F., Ganguly, N., Gummadi, K.P.: Understanding and Combating Link Farming in the Twitter Social Network. In: Proc. of the 21st int. conference on World Wide Web, WWW '12, pp. 61–70 (2012)
14. Guimerà, R., Amaral, L.: Cartography of complex networks: modules and universal roles. J Stat Mech **2005**(P02001), nihpa35,573 (2005)
15. Java, A., Song, X., Finin, T., Tseng, B.: Why we twitter: understanding microblogging usage and communities. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, WebKDD/SNA-KDD '07, pp. 56–65 (2007)
16. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proc. of the 19th int. conference on World wide web, WWW '10, pp. 591–600 (2010)
17. Lakshman, A., Malik, P.: Cassandra: a structured storage system on a p2p network. In: Proc. of the 28th ACM symp. on Princ. of distributed comput., PODC '09, pp. 5–5 (2009)
18. Martínez-Bazan, N., Águila Lorente, M.A., Muntés-Mulero, V., Dominguez-Sal, D., Gómez-Villamor, S., Larriba-Pey, J.L.: Efficient Graph Management Based On Bitmap Indices. In: Proc. of the 16th Int. Database Eng. & Appl. Symp., IDEAS '12, pp. 110–119 (2012)
19. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. In: Proceedings of the 8th international conference on Autonomic and trusted computing, ATC'11, pp. 175–186 (2011)
20. OrientDB: (1999). http://http://www.orientdb.org/
21. Rodgers, S.: Twitter blog (2013). https://blog.twitter.com/2013/behind-the-numbers-how-to-understand-big-moments-on-twitter
22. Schatz, M.C., Langmead, B., Salzberg, S.L.: Cloud computing and the DNA data race. Nat. Biotech **28**(7), 691–693 (2010)
23. Schuett, T., Pierre, G.: ConpaaS, an integrated cloud environment for big data. ERCIM News **2012**(89) (2012)

24. Simpson, G.G.: Mammals and the nature of continents. Am. J. of Science (241), 1–41 (1943)
25. Suh, B., Hong, L., Pirolli, P., Chi, E.H.: Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In: Social Computing (SocialCom), 2010 IEEE Second International Conference on, pp. 177–184 (2010)
26. Thomas, K., Grier, C., Song, D., Paxson, V.: Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, IMC '11, pp. 243–258 (2011)
27. Thusoo, A., Sarma, J.S., Jain, N., Shao, Z., Chakka, P., Zhang, N., Anthony, S., Liu, H., Murthy, R.: Hive - a petabyte scale data warehouse using hadoop. In: IEEE 26th Int. Conference on Data Eng., pp. 996 –1005 (2010)
28. Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., Wilkins, D.: A comparison of a graph database and a relational database: a data provenance perspective. In: Proc. of the 48th Annu. Southeast Reg. Conference, ACM SE '10, pp. 42:1–42:6 (2010)
29. Wang, A.H.: Don't follow me: Spam detection in twitter. In: Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, pp. 1–10 (2010)
30. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. Nature **393**(6684), 440–442 (1998)
31. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, WSDM '10, pp. 261–270 (2010)

A Storage and computations

In a previous work [11], we aimed to detect social capitalists by using high-level tools and consuming a few amount of computing resources, namely CPU and memory. We wanted to design an easily reproducible method, available to everyone with a single desktop server. We focused on databases and especially the NoSQL and graph-oriented ones. Indeed, it seems natural to use databases when trying to handle data, even if these ones are organized as a graph. Furthermore, our computations, based on vertices neighborhood comparisons are quite straightforward and thus could be processed by using a reasonable amount of computing resources easily. However, after having tried MYSQL, several NoSQL tools like Cassandra [17] and graph-oriented tools such as OrientDB [20], Neo4J [28], we observed that most of these tools did not fit with our requirements. Either we were not able to load our graph in a reasonnable amount of time (Neo4j, OrientDB), either the calculation was far too long (MYSQL, Cassandra). We hence finally chose Dex [18], a high-performance graph-oriented database which allowed us to process half of our graph in a few hours. Still, we were not able to compute our measures on the whole graph because of Dex built-in malfunctions which are still not solved.

In this paper, we wanted to process our measures on the whole graph to detect all the social capitalists and to confirm our threshold in order to be able to use it on other graphs. We thus had to make a compromise. We are still using a few amount of resources, only one processor and less than 40GB of RAM for the whole graph, less than with Dex. However, we are not using anymore a high-level method. We are now working with taylor-made programs in C++ representing the graph in memory as a sparse matrix, which leads to very efficient calculations. Indeed, processing the overlap and ratio indices on the whole graph is made in a few minutes with this method. Other softwares from the field like the community detection method designed by Blondel et al [7] use this way of processing the graph. This method has nevertheless a drawback to take into account. The identifiers of the vertices have to be a sequence of integer starting from 0. Thus, it is often needed to renumber the vertices of the graph. Actually, it was the case for all the datasets used in this paper. Furthermore, deleting vertices means renumbering the vertices again.

Basically, the way to store an undirected graph is quite simple. Two integers vectors are used. In the first one, the cumulative degrees of the vertices ranked according to their identifiers are stored. Its length is equal to the number of vertices. In the second one, the edges of the graphs are stored, represented by the list of neighbors of the vertices ranked by their identifiers. Its length is equal to twice the number of edges. The cumultative degrees of the first vector are used to find the neighbors of the vertices in the second vector. The neighbors of the vertex of identifier $i>0$ can be found in the second vector from the position $cumulative(i-1)$ to the position $cumulative(i)-1$. For instance, as one can see on Figure 9, the neighbors of the vertex of id 1 are elements from 1 ($cumulative(0)$) to 3 ($cumulative(1)-1$) in the second vector. In the case of a directed graph like in this paper, we have to use four vectors, two for the in-arcs and two for the out-arcs.

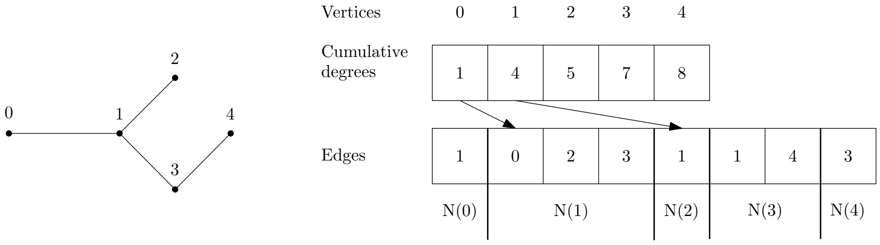


Fig. 9 A graph at left and his memory representation as a sparse graph.