HAL
open science

On the importance of considering social capitalism when measuring influence on Twitter

Maximilien Danisch, Nicolas Dugué, Anthony Perez

To cite this version:

Maximilien Danisch, Nicolas Dugué, Anthony Perez. On the importance of considering social capitalism when measuring influence on Twitter. BESC 2014 - International Conference on Behavioral, Economic, and Socio-Cultural Computing, Oct 2014, Shanghai, China. pp.1-7, 10.1109/BESC.2014.7059501 . hal-01105133

HAL Id: hal-01105133

https://hal.science/hal-01105133v1

Submitted on 19 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

On the importance of considering social capitalism when measuring influence on Twitter

Maximilien Danisch*†, Nicolas Dugu‡ and Anthony Perez‡
* Sorbonne Universits, UPMC Univ Paris 06, UMR 7606, LIP6, F-75005 Paris, France.
† CNRS, UMR 7606, LIP6, F-75005 Paris, France.
‡ Univ. Orlans, INSA Centre Val de Loire, LIFO EA 4022 45067, Orlans, France.

***Abstract*—Influence on Twitter has drawn a lot of attention these past few years since this microblogging service is used to share, seek or debate about any kind of information. Several tools providing so-called *influential scores* have thus been proposed. However, the algorithms behind them are kept secret and it is not clear how they consider influence. Yet, many users rely on such tools to evaluate and even try to improve their influence in the Twitter network. In a recent work, it has been shown that automatic accounts can obtain high influential scores with no intuitive reason. Extending and completing this work, we show that such measures fail at distinguishing so-called *social capitalists* from real, truthful users. This enlights the fact that actual scores do not seem to consider the way followers and interactions are obtained on the network. To overcome this issue, we define a classifier that discriminates social capitalists from truthful users. To that aim, we crawled the Twitter network to gather examples of certified social capitalists and regular users and obtained features related to the profile and behavior of each user. We then use such a classifier to balance Klout's score to adjust influential scores. We also developed an application that allows using our classifier online. We believe our work should raise the question of the legitimacy of influence on Twitter, and lead to significant improvements in the way it is measured.**

I. Introduction

Context. Twitter is a microblogging service that is mostly used to share, seek or debate about any kind of information [13]. The users of Twitter are allowed to post messages no longer than 140 characters, called *tweets*. This particular format allows users to share information in a very prompt manner, and there are about 1 billion tweets posted every two and a half day [17]. In the last few years, Twitter has also known an important growth in its number of users, going from 200 millions in April 2011 [5] to more than 500 millions in October 2012 [6]. This microblogging service can moreover be considered as a social network, since it includes several social features between users. There are mainly two types of such interactions. First of all, in order to see the tweets posted by a user, one has to *follow* his account. Following standard terminology, when user u follows user v we say that u is a *follower* of v and that v is a *friend* of u. Another important interaction between users is the notion of *retweet* [21] and *mention*. The first operation allows a user to share the tweet of another user to his own set of followers. Being retweeted is of important interest for Twitter users, since it provides a better visibility to their tweets. Regarding the second interaction, any user can *mention* another user in a tweet using the syntax @*Username*. This is for instance used to draw the attention of a particular user to a given tweet. These specific properties naturally lead to consider the notion of *influence* on the Twitter network, a topic that raised a lot of research recently [7], [9], [18], [22], [23]. There are a lot of parameters that can be taken into account when measuring influence on Twitter, mainly regarding the number of followers, retweets and mentions. Intuitively, the more a user is followed, retweeted and mentioned, the more his influence in the network should be considered important [9]. Notice that these parameters are usually not considered as equally relevant. Moreover, refined parameters can be considered, such that the *Friend/Follower*, *Retweet and Mention* and *Interactor* ratios [7]. Several tools have been proposed in order to associate to each user a so-called *influential score* that illustrates his influence in the network (see e.g. Klout [1], Kred [2], Tweet Grader [3], Twitalyzer [4]). In all cases, the underlying algorithm that provides a score for a user is kept secret. However, several insights are provided by Klout and Kred towards its process (see *e.g.* [7]). An important thing is that the number of followers is not a key parameter. Indeed, these tools focus instead on the importance of the interaction a user has with the network, Klout describing influence as *the ability to drive action*. Similarly, Kred measures two different parameters, namely *influence* and *outreach level*, and states the first one *goes up when somebody mentions, retweets or replies to you.*

Social capitalists. It has been recently observed that a lot of users are trying to gain as many followers as possible in an artificial manner [11], [12]. Roughly speaking, these users cheat with the friend-follower relationship by promising users that follow them to follow them back, or by following users regardless of their contents, just hoping to be followed back. The reason for this behavior is twofold. First, the more users are followed, the more they are visible on search engines of the network [12]. Moreover, their tweets are more likely to be retweeted, having a direct impact on their influential scores. Such users (so-called *social capitalists* [12]) use several means to achieve that goal, and this is sometimes the only activity they have on Twitter. It is important to notice that most of these accounts are real, active and even famous users [12]. They are also different from spammers since they are not spreading spam or malicious links. The efficiency of this method has been recently illustrated by Dugué and Perez [11], who created an automatic account (@*Rain_bow_ash*) that tweets on social capitalism-related matter exclusively. This account quickly gained a high number of followers and was highly retweeted despite the content of his tweets, only dedicated to call for new followers. Notice that such users encourage other users to retweet or mention them in order to gain many followers. They are thus very active on the network, but without any content-related reason. A natural question that follows these

observations is thus the following: *Should users that try to be followed by any means and asks explicitly for interaction be considered as influent as regular, truthful users?*

While it seems obvious that this question should be answered negatively, we will see that current tools measuring influence on Twitter do not make such a distinction.

Related work. Recently, Messias et al. [15] initiated a study providing some relevant information towards this problematic. They created two automatic accounts using a very simple strategy, and managed to grow one of them to obtain very high Twitalyzer and Klout scores. This bot automatically gained 500 followers, which is not much and tweeted about obvious popular topics automatically too. Thus, we cannot consider this account as influent and it is interesting to observe that tools like Twitalyzer and Klout stated the contrary. Besides, this account acts as a *cyborg*, namely an automatic account looking like a human. According to Chu et al. [10], these accounts can be efficiently detected using straightforward features like the source used to post tweets or the daily and weekly tweet schedule. However, both these works do not tackle the problem of measuring the influence of social capitalists. Indeed, as Dugué and Perez [11] stated, most of these users are real humans tweeting in a manual way, but gaining efficiently followers and being highly retweeted.

Contribution. In this paper, we follow this line of research and emphasize the fact that these measures do not distinguish social capitalists (Section II) from regular users. To highlight this, we describe a new Twitter dataset crawled based on hashtags related to social capitalism such as *#TeamFollowBack* (Section III). Hashtags are specific keywords used by Twitter users (beginning with #) to *tag* their messages and thus become more visible on search engines. By studying such users as well as certified social capitalists detected by Dugué and Perez [11], we observe that some of them are considered as highly influent by those measures (Section IV-A). To overcome this issue, we define a classifier that discriminates social capitalists from truthful users. Our algorithm relies on the notion of Logistic Regression, a classical Machine Learning tool. We next use the fact that our classifier outputs the probability for a given user to be a social capitalist to balance influential scores (Section V-B). We mainly focus on Klout's score because it is one of the most widespread tool. Finally, we present an online application that we developed to allow Twitter users to estimate the probability for a given user to be a social capitalist (Section V-C).

II. Social capitalists

Social capitalists are particular Twitter users that have drawn a lot of attention recently [11], [12], [15]. These users are trying to get as many followers as possible using two straightforward principles:

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

Such users are not healthy for a social network, since they promote users regardless of the content of their tweets, hence increasing the influence of malicious users such as spammers [12]. Based on the previous observations, Dugué and Perez [11] recently provided a method to detect social capitalists using two simple topological measures, namely *overlap index* [20] and the Friend/Follower ratio. The former computes the relationship that exists between the set A of friends and the set B of followers a user has, and is formally defined as:

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

Intuitively, a user with an overlap index close to 1 should be friend with a large majority of his followers, and thus have applied the aforementioned principles. On the contrary, a user with an overlap index close to 0 has a small intersection between his sets of friends and followers, and is thus less likely to be a social capitalist. Users that have an overlap close to 1 can next be distinguished between **FMIFY** and **IFYFM** users thanks to the Friend/Follower ratio, which compares the number of friends a user has to his number of followers. Indeed, a user with a ratio much greater (resp. smaller) than 1 is probably a **IFYFM** (resp. **FMIFY**) user.

Furthermore, Dugué and Perez [11] showed that social capitalists apply the principles mentioned above in an efficient way by tweeting on specific *hashtags* dedicated to social capitalism (see Figure 1). Social capitalists use this aspect of Twitter to interact with other social capitalists, for instance using hashtags such as *#IFollowBack* or *#TeamFollowBack*. Another specific behavior of social capitalists is that they explicitly ask for interaction (retweet or mention) in order to be followed (see Figure 1). Moreover, they encourage all users that retweet their tweets to follow each other, as illustrated by the use of the hashtag *#FollowTrain*. We make use of these easy-to-spot techniques to construct our dataset of social capitalists, as explained in the next Section.

Fig. 1. Timeline of the social capitalist *followback_707* .

III. Dataset

Positive examples. In order to constitute a dataset of certified social capitalists such as those described by Dugué and Perez [11], we gathered tweets posted on the specific hashtags *#TeamFollowBack*, *#instantfollowbackdedicated* and *#teamautofollow*. We identified the users that posted tweets with these hashtags several times in a small amount of time as social capitalists. We crawled a sample of roughly $25,000$ such users and consider it as our set of positive examples of social capitalists.

Negative examples. The first step to get negative examples was to randomly sample Twitter. Indeed, random users should

not be social capitalists. Twitter claims having more than 550 million users. We thus chose to get randomly $15,000$ users on the network. To that aim, we randomly picked $15,000$ integer identifiers between 0 and $550,000,000$. However, most of these users only have a few connections to the rest of the network and are thus not sufficient to constitute an exhaustive dataset of negative users. To get more connected and more active users on the network, we chose to randomly pick $55,000$ users among the friends of these users. All these users constitute our negative examples of social capitalists. We may wonder whether some of these randomly picked users are social capitalists. According to Dugué and Perez [11], this seems strongly improbable. Indeed, they detect a bit more than $160,000$ social capitalists in the dataset provided by Cha et al. [9] that contain 55 millions of users. Assuming that the number of social capitalists grows as fast as the whole network, we may consider that social capitalists form 0.2% of the network. We can reasonably consider that the probability of picking a social capitalist is uniform. Thus, we may consider that among the $55,000$ users considered as negative examples, 0.2% (namely 110) are actually social capitalists, which is negligible.

Then, by using the Twitter REST API, we crawled all these accounts to obtain relevant features allowing to characterize them. Such features can be classified in several categories (see Table I). We would like to mention that all the retrieved features can be easily obtained from the Twitter REST API. For instance, one can use the so-called methods *users/show* to get the classical parameters (number of friends, followers, tweets, listed, favorites) and *statuses/user_timeline* to obtain more refined features (such as the distinction between original tweets and retweets). All these information regarding tweets and users' profiles are provided in the JSON format by Twitter, which is thus quite easy to parse.

Notice also that due to Twitter's restrictions for the use of their API, the statistics we retrieve only consider the 200 last tweets sent by the user at hand.

Some of these features such as the sources used to post tweets and the number of url per tweet were proved to be discriminant to separate human accounts from automatized accounts [10]. However, Dugué and Perez [11] showed that most of the social capitalists tweeting on social capitalists hashtags are not using tools to automate their accounts. We thus need more features in order to provide an efficient classifier.

One can see that the features chosen seem relevant to discriminate social capitalists from truthful user (Figure 2). Indeed, we see that the distribution of the features is different for truthful users and social capitalists. It is also particularly interesting to see that social capitalists are more retweeted than regular users, the number of retweet being a feature at the heart of Klout and Kred scores.

IV. Measuring influence on Twitter

Determining the *influence* of a user on Twitter is a matter that raised a lot of research in the past few years [7], [9], [18], [22], [23]. There are a lot of parameters than can be taken into account in order to measure influence: number of followers, retweets, mentions, favorites or lists for instance.

CATEGORY	FEATURES	
activity	number of:	
	1	statuses (*i.e.* tweets)
	2	lists containing the user
	3	tweets being favorited
local topology	number of:	
	4	friends
	5	followers
	6	users that are both friends and followers
tweets' content	average number of:	
	7	characters per tweets
	8	hashtags per tweets
	9	url per tweets
	10	mentions per tweets
tweets' characteristics		
	11	average number of retweets for a tweet
	12	average number of retweets for a retweet
	13	percentage of retweets among tweets
sources	using proportion of	
	14	Twitter official web application
	15	management dashboard tool
	16	automatic Follow or Unfollow tool
	17	automatic tweet tool
	18	other applications (Vine, Wiki, Soundcloud...)
	19	phones or devices applications

TABLE I. Description of the different features considered.

In order to produce more relevant influential scores, several classical *ratios* are considered rather than the aforementioned simple parameters. The simplest and most intuitive one is the *Friend/Follower ratio*. The smaller the result, the more people are interested in the content provided by the user. On the contrary, if the result is greater than 1, the user is likely to be considered as a *mass-follower*. This ratio can quickly lead to misinterpretations, and is thus often associated to other parameters related to the degree of *interaction* a user has on Twitter. Two other ratios can thus be considered to discriminate more precisely influence on Twitter, namely *Retweet and Mention ratio* and *Interactor ratio* [7]. The former one counts the number of tweets that are retweeted or lead to a conversation, divided by the total number of tweets. The latter one considers the number of *individual* users who retweet or mention the user divided by his number of followers. Since Twitter is primarily content-oriented but also relies deeply on personal relations, it seems relevant to consider both ratios at the same time. Based on this observation, Anger and Kittl [7] defined the *Social Networking Potential* of a user as the mean of the two aforementioned ratios.

In this work, we mainly focus on tools that are available online since they are used by Twitter users to measure (and even try to improve) their influence on the network. Notice that such tools are becoming more and more popular[1]. This is in particular the case for Klout [1], Kred [2] or Twitalyzer [4], even though the latter one stopped its activity in September 2013 (notice that the tool is still available online though). We would like to mention that Klout does more than measuring influence on Twitter, since it can use several famous social networks (for instance Facebook, Linkedin and Instagram) to provide its influential score. It is however possible to know which networks are used in order to obtain the Klout score,

[1]Klout has been recently bought by Lithium Technologies for 200 million dollars [19].

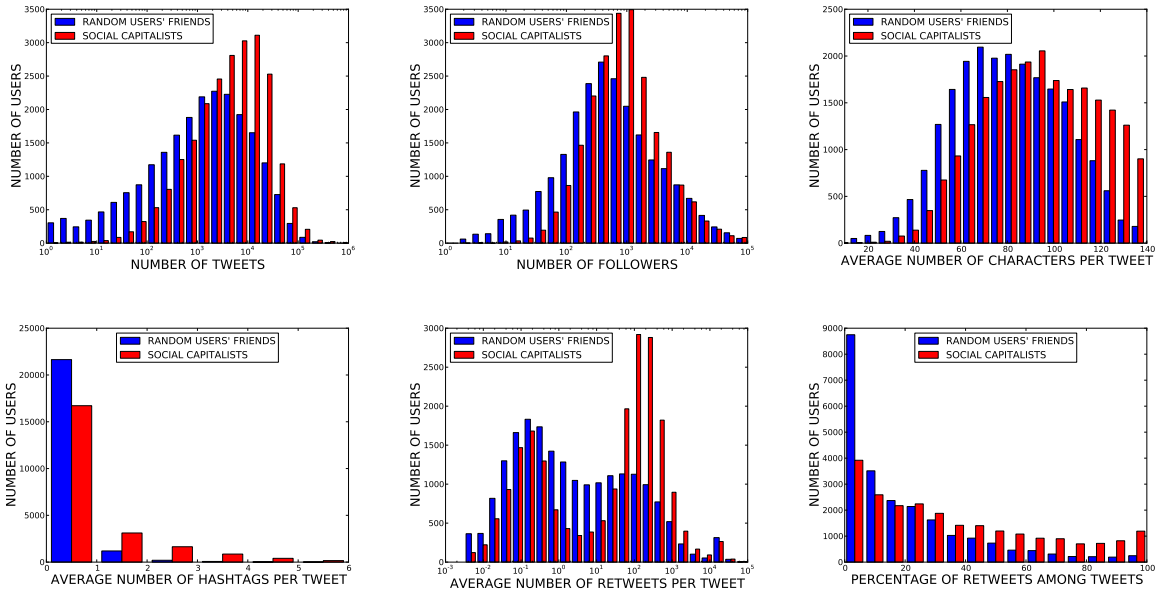

Fig. 2. Histograms showing statistics on some of the features for both, random users' friends and social capitalists.

and we will precise this every time we use such a measure. In the next section we will emphasize the limitations of such measures by presenting users that are obvious social capitalists but still manage to achieve high influential scores.

A. *The impact of social capitalism*

With the promess to follow back users that follow them, social capitalists gain many followers in a very short time. However, as mentioned previously, this should not ultimately imply an increase in their influential score, since this parameter is not highly considered by current tools. Nevertheless, having a large number of followers can obviously ease interactions such as retweets or mentions, and in turn increase influence. Moreover, social capitalists use another technique, that consists in asking for retweet and/or mention in order to follow or be followed by other users (see Figure 1). This behavior leads to a high level of interaction for these users, which in turn can also imply high influential scores.

We illustrate this fact by computing the Klout, Kred and Twitalyzer scores for certified social capitalists, either extracted from the dataset we collected, detected using the method provided by Dugué and Perez [11] or simply found on Twitter because their biography or even screen name contains social capitalism related sentences.

Name	Friends	Followers	Overlap	Klout	Kred	Twitalyzer
teamukfollowbac	120,065	134,669	0.99	79	98.9	25.8
berge31	2,522	2,434	0.97	76	77,8	1
TheDrugTribe	26,266	28,832	0.99	69	98.2	27.2
globalsocialm2	5,603	5,624	0.81	69	95.1	3.3
repentedhipster	3,148	2,940	0.98	66	78,2	1
LIGHTWorkersi	112,963	103,475	0.99	66	96,2	22.4
ilovepurple__	49,666	52,448	0.97	65	97.5	22.9
TEAMF0LL0W	13,246	78,615	0.97	65	99.2	21.2
TEEMF0LL0W	10,977	92,412	0.97	64	99.3	21.1

TABLE II. INFLUENTIAL SCORES OF SOCIAL CAPITALISTS EXTRACTED FROM OUR DATASET: KLOUT, KRED AND TWITALYZER.

Table II shows the different influential scores from several positive examples gathered for the sake of our dataset. Recall that such users have been crawled because they tweeted on social capitalism related content at some point. Along with this information, we also display the number of friends and followers as well as the overlap index users had at the time. Dugué and Perez [11] observed that any user with an overlap greater than 0.74 is likely to be a social capitalist, which is verified here. Those social capitalists are considered as influent users by all three measures. Some users even have explicit names or biographies, which do not however moderate their influence.

Name	Friends	Followers	Klout	Kred	Twitalyzer
EcheMadubuike	720,407	732,176	69	99	27.2
LarryWentz	600,196	660,260	60	90.9	20.2
machavelli7	567,553	572,161	68	94.8	20.9
zuandoemkta	566,971	555,476	60	94.4	20.6
_Follow_Friends	511,818	540,783	56	94.9	23.2
kosma003	438,050	423,638	52	95.1	20.2
ceebee308	360,163	382,568	60	98.1	22.6
ClimaWorld	384,365	375,419	60	90	20.4
radiotabu	306,066	336,752	55	97.7	21.5
Nteratedetodo	300,278	323,236	56	95.6	26.7

TABLE III. INFLUENTIAL SCORES OF SOCIAL CAPITALISTS DETECTED BY DUGUÉ AND PEREZ [11]: KLOUT, KRED AND TWITALYZER.

Table III presents similar results for social capitalists detected using the method of Dugué and Perez [11]. In all cases, the Klout score is taken from Twitter activity only. This explains the fact they are significantly smaller than the ones presented in Table II. Nevertheless, they remain by far above average. Moreover, one can see that their Kred and Twitalyzer scores are relatively high, once again implying that their activity is considered as influent. We would like to mention that the tweets of these users are not necessarily all related to social capitalism, and thus that considering their local network topology seems crucial in order to detect them. This is a real issue when measuring influence on Twitter.

To complete the previous observations, we compare in Figure 3 the different scores of two famous, active and popular Twitter accounts, namely **Barack Obama** and **Oprah Winfrey**, to certified social capitalists or automatic accounts (*Carina Santos*, the account created by Messias et al. [15]).

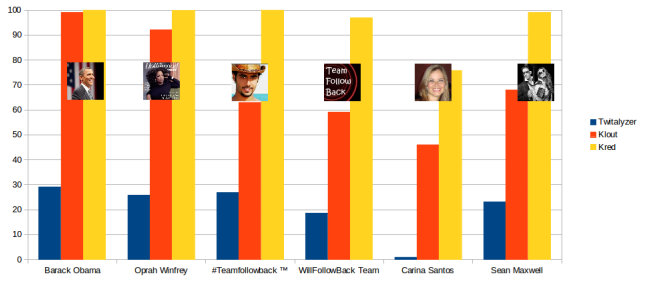

Fig. 3. Comparison of the three measures for the accounts of **Barack Obama** and **Oprah Winfrey** and obvious soical capitalists.

This comparison highlights several aspects of the different measures. First of all, one can observe that all social capitalists have similar Kred and Twitalyzer scores than the two accounts we use as references, while their Klout score is smaller. Actually, this difference can be easily explained: as mentioned previously, Klout computes a score from the activity of several networks, including **Wikipedia**. Both **Barack Obama** and **Oprah Winfrey** have a well-documented entry in **Wikipedia**, while the other accounts we consider are *exclusively* active on Twitter. Having a high Klout without combining several networks seem rather hard to achieve.

Except Carina Santos, the accounts we consider are obvious social capitalists that have explicit names and biographies on the network. None of the measures is able to detect that all accounts are automatic accounts, either bot or cyborg (that is automatic accounts that behave like human beings, see [10]). The accounts other than Carina Santos tweet *exclusively* about social capitalism and engage with other users with the unique objective to gain more followers. Despite the fact they do not produce any relevant content and instead post periodically and automatically the same kind of tweets, they are still considered as influent by all the measures.

To conclude this Section, we would like to notice that the average Klout score of social capitalists (taken from our dataset) grows linearly with their number of followers. Recall that Klout considers more the number of retweets and mentions than the number of followers. However, as mentioned previously, this observation emphasizes that it seems easier to have interactions with other users when the number of followers is large.

We believe that this lack of consideration of social capitalism drives malicious behavior on Twitter. Indeed, a lot of users are actively trying to improve their influential scores, which can be increased using techniques related to social capitalism. It is thus of important interest to take into account these parameters when measuring influence on Twitter.

V. A NOVEL APPROACH TO MEASURE INFLUENCE

We now present the main results of our work, namely a *classifier* that discriminates social capitalists (*positive examples)* from regular users (*negative examples*). The machinery we use (so-called logistic regression) allows us to obtain moreover the probability for a given user to be a social capitalist. As a first step towards providing a new measure of influence on Twitter, we make use of such a probability to balance actual influential scores (see Equation 1). As we shall see Section V-B, such a balancing measure allows us to lower efficiently the influence of social capitalists.

A. Classification of social capitalists

In a recent work, Dugué and Perez [11] provided algorithms that achieve a quite good accuracy for discriminating social capitalists from regular users. However, their aim was to rely on topological features *only*. In particular, this means they did not consider the content-related features of the tweets sent by users. Using this algorithm, they observed that social capitalists roughly represent 0.2% of the Twitter network, showing that detecting these users is non-trivial. While this is interesting from a theoretical perspective, we show these algorithms can be improved by considering additional relevant features. We thus use the dataset described in section III containing 77,102 users (22,845 social capitalists and 54,257 truthful users), each user being represented by the features described Table I. We splitted the dataset randomly into a 70% training set and 30% test set, the former one being used to adjust the classifier while the latter one is used to verify its accuracy[2] . We used classical Machine Learning classifiers, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR) [8]. We implemented these algorithms using the sklearn python library [16].

We obtained high and similar accuracieson the test set with all these classifiers, RF leading to slightly better results and KNN leading to slightly worse ones, similarly to the results obtained in [14]. Since it is specifically designed to obtain a probability to be a positive example (a feature that we will use to balance Klout's score), we decided to choose LR for our classifier. Another appealing characteristic of LR is that it is fast and easy-to-use after fitting for an online application (see Section V-C). Indeed, it only requires the storage of a single parameter for each feature.

In order to achieve a greater accuracy, and as it is classically done with the LR classifier, we transformed the features as follows: (i) we used an additionnal constant feature; (ii) since most of the features are more relevant in logarithmic scale, we used the logarithm of the features as input to the LR classifier; and (iii) we used their products two-by-two as additional features inputed to the LR classifier.

After fitting the parameters of the classifier that best classify the examples of the training set, we evaluated its performance on the test set. Figure 4 shows the histogram of the predicted probability of being a social capitalist. As one can see there is a very high correlation between our predicted probability and the fact that the example is a social capitalist or the friend of a random user. In addition, by cutting at a probability of 0.5, the classifier leads to an accuracy and an

[2]Accuracy is the proportion of correctly labeled examples.

F-score[3] of 87%. This means that by knowing only the few features detailed in Table I, we can efficiently predict whether a user is a social capitalist.

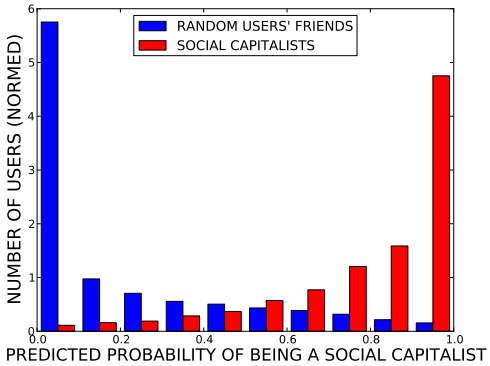

Fig. 4. Histogram of the predicted probabilities of being a social capitalist for random users' friends (i.e. truthful users) and users that posted on the specific hashtags (i.e. social capitalists).

In order to characterize more precisely the contribution of each feature, we also evaluated their relevance independently: we found that by using a single feature we hardly could reach an F-score of 60% (that is a bit better than a random classifier). This means that using a single feature is not enough to classify the users. In particular, when spammers can be discriminated efficiently using the number of urls [14], it is not the same with social capitalists which have a different tweeting behavior: most of their tweets are not dedicated to social capitalism (except for a few users) and their aim is not to spread spam.

We also evaluated the relevance of the 5 groups of features (see table IV). The more relevant are the one concerning the activity the users have on Twitter.

Feature	Accuracy	Sensitivy	Specificity	F-score
all features	86.0%	88.8%	84.9%	86.8%
activity	71.7%	82.8%	67.0%	74.1%
local topology	64.9%	73.4%	61.2%	66.8%
tweets' content	69.4%	63.9%	71.7%	67.5%
tweets' caracteristics	65.8%	68.0%	64.8%	66.4%
sources	67.3%	76.1%	63.6%	69.3%

TABLE IV. RESULTS OBTAINED ON THE TEST SET USING GROUPS OF FEATURES.

B. Balancing Klout's score

This LR classifier also gives us an evaluation of the probability for a user of interest to be a social capitalist, namely P_{Ksoc}. We can now use the quantity P_{Ksoc} to disadvantage the Klout score S_{Klout} (or any other score) and obtain a score S_{DDP} that takes into account social capitalism. As mentioned previously, we focus on Klout score because it is one of the most widespread influence measure for the Twitter network.

$$S_{DDP} = \begin{cases} S_{Klout} & \text{if } P_{Ksoc} \leq 0.5 \\ 2(1 - P_{Ksoc})S_{Klout} & \text{if } P_{Ksoc} > 0.5 \end{cases} \quad (1)$$

This score penalizes users that are classified as social capitalist by our classifier. We evaluate this new score on known social capitalists, influential users and random users taken from our dataset as well as previously observed accounts (see Table V). As one could expect, our classifier does not consider the two

Name	Klout score	P_{Ksoc}	S_{DDP}
barackobama	99	$8.42 \cdot 10^{-4}$	99
oprah	93	$5.86 \cdot 10^{-9}$	93
followback_707	64	0.999	1
seanmaxwell	69	0.937	9
scarina91	46	0.110	46
teamukfollowbac	80	0.838	26
TEEMF0LLOW	69	0.416	69
zuandoemkta	62	0.861	18
kosma003	53	0.747	27
NicolasDugue	33	0.120	33

TABLE V. KLOUT VERSUS DDP SCORES ALONG WITH P_{Ksoc}.

real and popular accounts we considered in this work (namely **Barack Obama** and **Oprah Winfrey**) as social capitalists. We would like to notice that the detection method of Dugué and Perez [11] does consider Barack Obama as a so-called *passive* social capitalist, that is an user that used to apply social capitalism techniques but has stopped a long time ago. This is thus consistent with our features, that consider the 200 last tweets of the user only.

When considering accounts from our positive examples test set (such as **teamukfollowbac**), one can observe that the probability of being a social capitalist is higher. Thus, the balanced Klout's score reflects this observation, and describes these users as less influent. This is even stronger when considering **seanmaxwell** and **followback_707**, two users also extracted from our positive examples test set. As one can see in Figure 1, the latter user only tweets in order to reach other social capitalists, a behavior that should not lead to a high influence on Twitter. The balanced score takes these considerations into account, and does not grant a high influence to such users. Finally, **TEEMF0LL0W** is an example of users where the classifier results are not completely convincing. This user which is a positive example often tweets very short messages with only one hashtag, which is not characteristic of a social capitalist. It is thus difficult to classify it correctly.

C. Online application

To complete our work we designed an online application that allows to compute the probability of being a social capitalist for any user. This application is available at the following URL: http://www.bit.ly/DDPapp (notice that one needs to own a Twitter account in order to use our application).

Given the screen name of a Twitter user, the application computes some of the aforementioned features and then makes use of the Logistic Regression classifier to compute the probability that this user is a social capitalist. The features are retrieved amongst the 200 most recent tweets of the given user. This restriction comes from the limit set by Twitter's API, that we cannot overcome at the moment. On the top

[3]F-score is the harmonic mean of sensitivity and specificity, sensitivity being the proportion of correctly labeled positive examples, and specificity the proportion of correctly labeled negative examples. It is better than the accuracy to quantify the performance of a binary classifier in case of non-equilibrated classes.

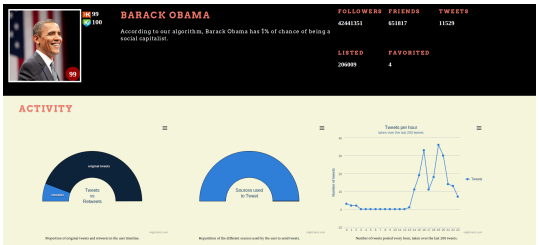

Fig. 5. Screenshot of the online application.

left of the page, Klout and Kred scores are displayed. We show the balanced Klout's score on the profile picture of the given user. Besides displaying the basic information regarding a Twitter account (number of friends, followers, tweets, lists and favorites), our application also displays some charts (see Figure 5). From left to right, the charts respectively represent the proportion of original tweets and retweets among the 200 last tweets, the different sources used to post tweets (*e.g.* the official web or smartphone applications, some automatic tools) and finally the hours when tweets are sent by users.

In order to provide a responsive and easy-to-use application, we also had to make some compromise regarding the use of our classifier. The main modification relies in the absence of Feature 6 which corresponds to the *overlap index* of a given user. The reason for this is simple: since this feature requires to obtain both the set of friends and followers of the user, it is very long to retrieve due to Twitter's API restrictions. However, we observed that lacking this feature in our classifier results in a really small loss of accuracy, and we thus chose to ignore it in the design of our application.

VI. Conclusion

In this work we confronted several tools measuring influence on Twitter [1], [2], [4] to recent works conducted on so-called social capitalists [11], [12]. Our study shows that a lot of users that do not tweet on any topic other than social capitalism are considered as influent. Indeed current measures do not seem to consider the impact of social capitalism on the Twitter network. To overcome these limitations, we provide a classifier that use overlap and Friend/Follower ratio indices as well as new parameters crawled on Twitter to detect efficiently social capitalists. This results extends a previous work [11], who detected social capitalists using the topology of the underlying relation-between-users graph only. We then use this classifier to balance Klout's score, which is one of the most popular tool measuring influence on Twitter. This allows us to lower the influential score of users that are obvious social capitalists while maintaining the score of regular users. Finally, we provide an online application to estimate the probability of being a social capitalist for any Twitter user.

This work can be extended in several manners. First of all, due to the Twitter rate limitations, our dataset is quite small. By getting a bigger dataset, we would be able to build an even more robust classifier. Furthermore, in this paper, we focus on distinguishing social capitalists from other users, considered as regular. We then used the results to balance the Klout score. It would be interesting to build a new influence score from scratch using efficient features. Besides, this work enlights the importance of considering new features. Indeed, because of the correlation between the number of retweets and the number of followers, the number of retweets alone cannot be considered as a good indicator of influence. A possible outcome for this would be to consider as more important retweets from users who are not followers of a user. The same remark can be made for mentions. Finally, this work is mainly focused on Twitter where social capitalists were introduced by Ghosh et al. [12]. Other applications with social features such as Instagram or Youtube may exhibit the same properties.

References

[1] Klout, the standard for influence. http://www.klout.com.

[2] Kred story. http://www.kred.com.

[3] Review your Twitter account — free Twitter analyzer — Twitter grader - http://twittergrader.mokumax.com/. http://twittergrader.mokumax.com/.

[4] Twitalyzer, serious analytics for social business. http://www.twitalyzer.com - As of September 28, 2013 Twitalyzer has decided to no longer sell new subscriptions.

[5] Twitter: We Now Have Over 200 Million Accounts, 2011. http://www.huffingtonpost.com/2011/04/28/twitter-number-of-users_n_855177.html.

[6] The Telegraph : Twitter in numbers, 2013. http://www.telegraph.co.uk/technology/twitter/9945505/Twitter-in-numbers.html.

[7] I. Anger and C. Kittl. Measuring influence on Twitter. In *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*, pages 1–4. ACM, 2011.

[8] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[9] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of ICWSM*. AAAI, 2010.

[10] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on Twitter: Human, Bot, or Cyborg? In *Proceedings of the 26th Annual Computer Security Applications Conference*, pages 21–30. ACM, 2010.

[11] Nicolas Dugué and Anthony Perez. Social capitalists on Twitter: detection, evolution and behavioral analysis. *Social Network Analysis and Mining*, 4(1):1–15, 2014. Springer.

[12] S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. Gummadi. Understanding and combating link farming in the Twitter social network. In *WWW*, pages 61–70, 2012.

[13] A. Java, X. Song, T. Finin, and B. Tseng. 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, pages 56–65, 2007.

[14] M McCord and M Chuah. Spam detection on twitter using traditional classifiers. In *Autonomic and Trusted Computing*, pages 175–186. Springer, 2011.

[15] J. Messias, L. Schmidt, R. Oliveira, and F. Benevenuto. You followed my bot! transforming robots into influential users in Twitter. *First Monday*, 18(7), 2013.

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[17] Simon Rodgers. Twitter blog, august 2013. https://blog.twitter.com/2013/behind-the-numbers-how-to-understand-big-moments-on-twitter.

[18] A. Sameh. A Twitter analytic tool to measure opinion, influence and trust. *Journal of Industrial and Intelligent Information*, 1(1):37–45, 2013.

[19] Catherine Shu. Tech Crunch, march 2014.

[20] George Gaylord Simpson. Mammals and the nature of continents. *Am. J. of Science*, (241):1–41, 1943.

[21] B. Suh, Lichan H., P. Pirolli, and E. H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in Twitter network. In *SocialCom*, pages 177–184, 2010.

[22] R. Tinati, L. Carr, W. Hall, and J. Bentwood. Identifying communicator roles in Twitter. In *Proceedings of the 21st International Conference Companion on World Wide Web*, pages 1161–1168. ACM, 2012.

[23] B. Waugh, M. Abdipanah, O. Hashemi, S. A. Rahman, and D. M. Cook". The influence and deception of Twitter: The authenticity of the narrative and slacktivism in the australian electoral process. In *Proceedings of the 14th Australian Information Warfare Conference*, 2013.