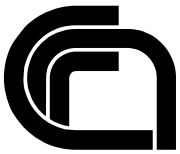

Consiglio Nazionale delle Ricerche

Fake accounts detection on Twitter

R. Di Pietro, S. Cresci, M. Petrocchi, A. Spognardi, M. Tesconi

IIT TR-15/2013

Technical report

Settembre 2013

Istituto di Informatica e Telematica

Fake accounts detection on Twitter

A taxonomy and some open issues

Roberto Di Pietro
IIT-CNR, Pisa, Italy
Dept. of Maths and Physics,
University of Roma Tre, Italy
dipietro@mat.uniroma3.it

Stefano Cresci, Marinella Petrocchi,
Angelo Spognardi, Maurizio Tesconi
IIT-CNR
Pisa, Italy
name.surname@iit.cnr.it

ABSTRACT

Fake followers are those Twitter accounts created to inflate the number of followers of a target account. Fake followers are dangerous to the social platform and beyond, since they may alter concepts like popularity and influence in the Twittersphere—hence impacting on economy, politics, and Society. In this paper, we provide several contributions. First, we review the most relevant existing criteria (proposed by Academia and Media) for anomalous Twitter accounts detection, and later we assess their capability to detect fake followers. In particular, we contribute with the creation of a gold standard of verified human, as well as with a set of known fake accounts. We test the above cited criteria against these two data sets, showing that the analyzed mechanisms provide unsatisfactory performance in revealing fake followers. Moreover, building upon these results, we also introduce a novel taxonomy to discriminate fake followers from legitimate ones and spammers. The findings reported in this paper, other than being supported by a thorough experimental methodology and being interesting on their own, also pave the way for further investigation.

Categories and Subject Descriptors

[**Security and privacy – Social network security and privacy**]

Keywords

Twitter, fake followers detection, gold standard, taxonomy

1. INTRODUCTION

Originally started as a personal microblogging site, Twitter has been transformed by common use to an information publishing venue. As of March, 2013, statistics reported 500 million of Twitter subscribers, with some 170 billion ($> 2^{39}$) of tweets sent [15]. Twitter annual advertising revenue in 2013 has been estimated to $399,500.000 [18]. Popular public characters, such as actors and singers, as well as traditional mass media (radio, tv, and newspapers) use Twitter as a new media channel. Politicians commit a notable part of their campaigning to their Twitter home pages, as it happened for the last US presidential and Italian general election events [22]. As a consequence, the Twitter platform has raised the attention of Industry and Business as well, with some (if not all) of the most famous brands massively using this platform for business promotion [3].

Such a versatility and spread of use have made Twitter the ideal arena for proliferation of anomalous accounts, that behave in unconventional ways. Academia has focused its attention on *spammers*, that is those accounts actively putting their efforts in spreading malware, sending spam and advertising activities of doubtful legality [25, 9, 21, 5]. Very often, to enhance the effectiveness of spammers, they are armed with automated twitting programs, known as bots. However, note that such automated pieces of software could be designed and used to post legitimate tweets as well—such as news updates.

Recently, media have started reporting that the accounts of politicians, celebrities, and popular brands featured a suspicious inflation of followers [6, 12, 7]. So called *fake followers* correspond to Twitter accounts specifically exploited to increase the number of followers of a target account. As an example, during the last 2012 US election campaign, the Twitter account of challenger Romney experienced a sudden jump in the number of followers. The great majority of them has been later claimed to be fake [12]. Similarly, before the last general elections in Italy took place (on the 25th of February 2013), online blogs and newspapers had reported statistical data over a supposed percentage of fakes of major candidates [23].

At a first glance, acquiring fake followers could seem a practice limited to foster one's vanity—a maybe questionable, but harmless practice. However, deepening the analysis reveals that artificially inflating the number of followers can also be finalized to make an account more trustworthy and influential, in order to stand from the crowd and to attract other genuine followers. Indeed, the more the supposed influence, the more those accounts with lots of followers will likely interfere with

the genuine followers. Similarly, if the practice were adopted by spammers, it could act as a way to post more authoritative messages and launch more effective advertising campaigns. The outcome could be the alteration of the concepts of popularity and influence in the Twittersphere, leading to formation of fictitious public opinion and possible impact on real world economy and Society. That is why fake followers detection is an issue worth addressing.

Fake followers detection seems to be an easy task for many bloggers, that suggest their "golden rules" and provide a series of criteria, to be used as red flags to classify a twitter account behavior. However, such rules are usually paired neither with analytic algorithms to aggregate them, nor with validation mechanisms. As for Academia, researchers have focused mainly on spam and bot detection, with brilliant results characterizing Twitter accounts based on their (non)-human features. To the best of our knowledge, however, there is a lack of analysis on fake followers characterization and detection. Moreover, most of the scientific studies generates a classifier to discriminate twitter accounts. The classifier is built as follows: researchers manually test the nature of a set of accounts, that, upon testing, becomes the training set for a machine learning-based classifier. The intuitive drawback is that humans are not error-free, and, thus, the manual classification phase is both error prone and time consuming. This latter drawback severely limiting the size of the data set taken used for the training.

Contributions

The goal of this work is to shed light on the phenomenon of fake followers, aiming at overcoming current limitations in their classification and detection. In particular, we provide several contributions. First, we provide a reference set of Twitter accounts, a so-called *gold standard*, where humans and fakes are known a priori. Second, we test several proposed methodologies for bot and spam detection on our gold standard. The outcome of the analysis leads us to assert that fake detection is a different task than that of spambot detection. Fake followers deserve other specialized mechanisms. Last, but not least, we provide a taxonomy depicting the main differences between spambots, humans, and fake followers, using several features as dimensional spaces to accurately classify where the accounts lie.

Organization

The remainder of this paper is organized as follows. Section 2 considers related work in the area of Twitter spam and bot detection. Section 3 describes our reference dataset. In Section 4, we concentrate on a set of criteria for anomalous account detection. The outcome of the application of the criteria over our reference dataset is described in Section 5. Section 6 depicts a novel taxonomy of Twitter followers. Finally, Section 7 concludes the paper. An Appendix is provided to detail the detection criteria.

2. RELATED WORK

In this section, we revise recent work in the area of spam and automation detection of user behavior on Twitter.

The work in [21] presents an analysis on how spammers operate on Facebook, Twitter, and MySpace. For data gathering, the authors created a large set of honey profiles on the three social platforms, logged the kind of contacts and messages that they received, and manually analyzed the collected data. The analysis reported that the suspicious accounts shared some common traits, formalized by the authors in a set of features. They served as input to a machine learning based classifier [], to take automatic decisions over a large set of unknown accounts. Impressively, such an approach led to the detection of more than 15k spam profiles, that Twitter promptly deleted.

In [25], the authors observe that the more researchers and engineers make progress in keeping Twitter a spam-free online community, the more Twitter spammers are evolving to evade existing detection techniques. They also propose a taxonomy of criteria for detecting Twitter spammers. A series of experiments show how the newly designed criteria feature a detection rate higher than existing ones.

Authors of [5] classify Twitter accounts in three classes: humans, bot, and cyborgs. The latter class represents either a bot-assisted humans or an human-assisted bots. Six thousands accounts have been manually classified to create a training set and a test set, each one with 1000 accounts for each of the three classes. The authors build their classifier based on four components: an entropy component that evaluates the timing regularity of an account tweets; a spam filter to detect spam tweets; an account property analyzer to extract additional information; and a decision maker component. This last one determines the class of a given account combining the outputs of the other three parts with a multiclass linear discriminant (LDA) analysis method.

Work in [20] makes an interesting analysis on the underground phenomenon of so called Twitter Account Markets, i.e., websites offering their subscribers to provide followers in exchange for a fee, and to spread promotional tweets on their behalf. The authors list a series of criteria that are helpful to detect Account Markets clients that pay for acquiring followers and spam with debatable tweets. In addition, criteria for detecting the spammer victims are also highlighted. Results of the analysis reveal a surprising and alarming business behind this phenomenon.

A series of reports (available at *digitalevaluations.com*) have attracted the attention of Italian and European newspapers and magazines (see, e.g., [4]), raising doubts on the Twitter popularity of politicians and leading international companies. A number of criteria, inspired by common sense and denoting *human* behavior, are listed in the reports and used to evaluate a sampling of the followers of selected accounts. For each criterion satisfied by a follower, a *human* score is assigned. For each not fulfilled criterion, either a *bot* or *neutral* score is assigned to the account. According to the total score achieved, Twitter followers are classified either as humans, as bots or as neutral (in this last case, there is no sufficient information to assess their nature), providing a quality score of the effective influence of the followed account. The results in [4] lack a validation phase.

Beside academic work, we assisted to the proliferation of online blogger and columnist posts, listing their own criteria for Twitter bots detection. As an example, a well-known blogger in [17] indicates as possible bots-like distinctive signals the fact that bots accounts: 1) have usually a huge amount of following and a small amount of followers; 2) tweet the same thing to everybody; and, 3) play the follow/unfollow game, meaning that they follow/un-follow an account usually within 24 hours. Criteria advertised by online blogs are mainly based on common sense and the authors usually do not even suggest how to validate them.

Finally, some companies specialized in social media analysis, like [19, 16], offer online services to analyze how much a Twitter account is *genuine* in terms of its followers. However, the criteria used for the analysis are just partially deducible from information available on their web sites.

In the following, we apply to our datasets the criteria for spam and bots detection proposed by five of the cited work, namely [21, 25, 4, 17, 16]. We are aware that this selection is not exhaustive. However, it considers a huge collection of criteria, that we further leverage for our reasoning on fake follower detection. It is worth noticing how other work for spam detection, like [9, 26], base their results on a subset, or on a slightly modified version, of the criteria considered by our selected set of works.

3. REFERENCE DATASETS

In this section, we introduce the datasets of Twitter accounts that will be used throughout the paper.

3.1 The Fake Project

The Fake Project started its activities on December 12, 2012, with the creation of the Twitter account @TheFakeProject. Its profile reports the following motto: *Follow me only if you are NOT a fake* and explains that the initiative is linked with a research project owned by researchers at IIT-CNR, in Pisa-Italy. The account biography points to the project web page *http://wafi.iit.cnr.it/TheFakeProject/*. At that page, one may find instructions to join the initiative and an overall description of motivations and goals of the project. In a first phase, the owners contacted further researchers and journalists to advertise the initiative. The online version of a popular Italian newspaper and a famous Italian social media analyst promoted the project and invited people to join it (see [14, 8] for an Italian version of these pieces). Foreign journalists and bloggers also supported the initiative in their countries. In a twelve days period (Dec 12-24, 2012), the account has been followed by 574 followers. Through Twitter API v1.1, we crawled a series of public information from these followers, i.e., their profiles and timeline information, together with their followers and followings profiles. In one day, we crawled these 574 accounts, leading to the collection of 616,193 tweets and 971,649 relationships.

All those followers voluntarily joined the project. To include them in our reference set of humans, we also launched a verification phase. Each follower received a direct message on Twitter from @TheFakeProject, containing an url to a CAPTCHA, unique for each follower. We consider as "certified human" any account that completed the CAPTCHA. We verified 469 out of the 574 followers.

3.2 #elezioni2013 dataset

The #elezioni2013 dataset was born to support a research initiative for a sociological study carried out in collaboration with the University of Perugia and the Sapienza University of Rome, on the strategic changes in the Italian political panorama for the 3-year period 2013-2015. Researchers identified 84,033 unique Twitter accounts that used the hashtag #elezioni2013 in their tweets, during the period between January 9 and February 28, 2013. Identification of these accounts has been based on specific keyword-driven queries on the username and biography fields of the accounts' profiles. Keywords include blogger, journalist, social media strategist, congressperson, representative. Specific names of political parties have been also searched. In conclusion, all the accounts belonging to politicians and candidates, parties, journalists, bloggers, specific associations and groups, and whoever somehow was officially involved in politics, have been discarded. Accounts not having a biography have been discarded too. The remaining accounts (about 40k) have been classified as *citizens*. This last set has been sampled (with confidence level 95% and confidence interval 2.5), leading to a final set of 1488 accounts, that have been subject to a manual verification to determine the nature of their profiles and tweets. Finally, 1481 accounts became

part of dataset #elezioni2013.

3.3 Gold standard of human accounts

The above introduced datasets form our final set of about 1950 verified human accounts. It is worth noticing how the two subsets differ from each other. The Fake Project consists of accounts that have been recruited on a volunteer base: people involved in the initiative aimed to be part of an academic study for discovering fake followers on Twitter, and are a mixture of researchers and social media experts and journalists, mostly from Italy, but also from US and other European countries. The #elezioni2013 set consists of particularly active Italian Twitter users, with different professional background and belonging to diverse social classes, sharing a common interest for politics, but that do not belong to the following categories: politicians, parties, journalists, bloggers.

3.4 Gold standard of fake followers

In April, 2013, we bought 3000 fake accounts from three different Twitter online markets. In particular, we bought 1000 fakes accounts from *http://fastfollowerz.com*, 1000 from *http://intertwitter.com*, and 1000 fake accounts from *http://twittertechnology.com*, at a price of $19, $14 and $13 respectively.

4. CRITERIA SPECIFICATION

In this section, we focus on 5 works available in the literature [4, 25, 21, 16, 17], also reported in Section 2. For each work, we report the authors' observations and criteria for anomalous account detection, further specifying how we converted them into a rule to be run over our datasets. We aim at testing the selected rules on the datasets, to assess their strength to discriminate fake followers. We remark that giving a judgment on the overall quality of such criteria is beyond the scope of this paper. They have been originally designed for spam detection and here, for the first time, we apply them to another category of Twitter accounts, i.e., the fake followers.

It is also worth noticing that we focus on the application of each single rule. Indeed, in many cases, the analyzed paper did not specify the algorithm for aggregating the proposed rules to have them to act as a detector. Details on how aggregation has been performed (and hence, it is realizable) are provided in [4] only. For this work, we present both the results of the single rules application and the overall aggregation results.

The analyzed criteria are divided into three main groups: those derived from academic work, those derived from blogs, and those derived from online specialized firms. Throughout the sequel of the paper we use the term "friends" to denote the users followed by an account (i.e., if A follows B, B is a friend of A).

4.1 Academic work

Among the works introduced in Section 2, we considered the most focused on automatized spam detection.

4.1.1 Detecting spammers in social networks

The work presented in [21] focuses on detecting spambots and exploits five features, that can be gathered crawling an account's details. Per each account, such features are given as input to a Random Forest algorithm, that outputs if the account is a spambot or not. Without the original training set, we were unable to reproduce the same classifier, but we picked its features, as summarized in Appendix A.1, and we report its detection effectiveness in Section 5.

4.1.2 Fighting evolving Twitter spammers

The authors of [25] observed that Twitter spammers often modify their behavior in order to evade existing spam detection techniques. Thus, they suggested new metrics, making evasion more difficult for spammers. Beyond the metrics directly available from the account profile lookup, the authors propose some graph-, automation-, and timing-based metrics, as detailed in Appendix A.2. The authors propose to combine their metrics in four different machine learning classifiers and compare their implementation with other existing approaches. We were unable to completely reproduce their machine learning classifiers since we had a different dataset, but here we evaluate most of those metrics claimed by the authors to be quite robust against evasion techniques.

4.1.3 Followers of political candidates

Camisani [4] carried out a series of tests over samples of Twitter followers of Romney and Obama, for the last US presidential elections candidates, as well as for popular Italian politicians. In particular, in [4] it is detailed an algorithm to evaluate the account nature based on some of its public features. For each feature, the algorithm assigns a *human* and a *bot/inactive* score and classifies an account considering the gap between the two scores. The algorithm reported in [4] has enough details to be fully implemented. We report in the following section the results of testing each single rule, as well as their aggregate.

4.2 Blogs

Several bloggers propose their golden rules to identify suspicious Twitter accounts. Here, we consider the *7 signals to look out for recognizing Twitter bots* (see Appendix A.4), according to the founder of the social media website *stateofsearch.com* [17]. As detailed in Appendix, we were able to apply only a subset of the seven suggestions to our gold standard and the results are in Table 1.

4.3 Online analyzers

Several social media firms provide online tools to classify Twitter followers based on a *fakeness* degree. Here, we consider the FakeFollowerCheck tool, by Socialbakers [16]. Their website provides eight criteria (detailed in Appendix A.5) to evaluate the fakeness degree of the followers of a certain account, but omits details on how to combine such criteria to classify the account. We contacted the customer service of the social media firm, but were answered that "how the respective criteria are measured is rather an internal information". In Table 1, we report how each single criterion is able to classify our accounts.

5. RESULTS AND DISCUSSION

All the rules detailed in Appendix A, and related to the work recalled in Section 4 have been applied to a mixed dataset, composed of a priori known human accounts, belonging to The Fake Project (469 verified accounts) and to #elezioni2013 (1481 verified accounts), as well as the fake accounts bought from the Twitter account markets. In particular, we randomly chose 1950 out of the 3000 fake accounts we have, to obtain a mixed and balanced dataset of 3900 account composed by 50% of humans and 50% of fakes.

5.1 Evaluation

Being interested in fake follower detection, we considered the ability of each rule to detect a fake account. The results of rules application are summarized in Table 1 and, specifically, in the table we reported:

- True Positive *TP*: the number of those fake followers recognized by the rule as fakes;
- True Negative *TN*: the number of those human followers recognized by the rule as humans;
- False Positive *FP*: the number of those human accounts recognized by the rule as fakes;
- False Negative *FN*: the number of those fake accounts recognized by the rule as humans.

To evaluate the quality of a rule, we consider the following standard evaluation metrics:

- Precision: the proportion of predicted positive cases that are indeed real positive, that is $\frac{TP}{TP+FP}$;
- Recall: the proportion of real positive cases that are predicted positive, that is $\frac{TP}{TP+FN}$;
- *F-Measure*: the harmonic mean of precision and recall, namely $2\frac{precision \cdot recall}{precision+recall}$
- Matthew Correlation Coefficient *MCC* [2]: the estimator of the correlation between the predicted class and the real class of the samples. This metric is considered the unbiased version of the F-Measure, since it uses all four elements of the confusion matrix:

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FN)(TP+FP)(TN+FP)(TN+FN)}}$$

Since an MCC value close to 1 means that the prediction is really accurate, in Table 1 we highlight those criteria whose application gives $MCC \geq 0.75$: among the fifty rules that we have applied to our mixed dataset, such rules have the strongest correlation with the typology of the accounts.

Finally, for such rules that generically say "profiles with high (resp., low) values of attribute X are less likely to be Y" (where, in the original rule, Y can be either spambot or human), we consider a fixed threshold evaluated according to the common rule of information gain with continuous attributes, see [11]: we sort the examples according to the continuous attribute X, we identify the possible threshold values, and evaluate the information gain associated with each of them, to select the best performing one. In the Appendix, we precisely report, rule by rule, the used thresholds.

Please notice that for some rows of the table the number of evaluated accounts is lower than 3900. This is because we were unable to evaluate some of the proposed ratios for many accounts of our dataset. For example, 1554 accounts had no tweets from API, then the API URL ratio was impossible to evaluate ($\frac{0}{0}$).

5.2 Discussion

In Table 1, we mark as highlighted those rules whose application on our mixed dataset obtained an *MCC* value ≤ 0.75. Visibly, only four rules passed this test, namely, the friends to followers$^{\frown 2}$ ratio, the bidirectional link ratio, the API ratio, and the threshold 30 over the number of followers. Even by lowering the *MCC* threshold, e.g., by fixing it equal to 0.6, only two further rules passed the threshold.

Noticeably, none of the criteria suggested by online blogs and those addressed by online tools, such as the SocialBakers FakeFollowerCheck, are successful in detecting the fakes in our dataset. Hence, this call for further investigations to characterize the account Twittersphere.

We acknowledge that our fake followers dataset could be illustrative, and not exhaustive, of all the possible existing sets of fakes. However, it is worth noticing that we found the Twitter accounts marketplaces by simply Web searching on the most common search engine. Thus, it can be argued that our dataset represents what is usually possible to be found on the Web.

Overall, we observe that the four rules in Table 1 that have obtained the best evaluation metrics, although applied in past work to detect spam behavior, are the best

		results				evaluation metrics			
	rule description	*TP*	*TN*	*FP*	*FN*	*precision*	*recall*	*F-M.*	*MCC*
Stringhini et al. [21]									
1	friends $\leq$ 500	1527	578	1372	423	0.527	0.783	0.629	0.091
2	tweets $\leq$ 20	717	1882	68	1233	0.913	0.368	0.524	0.415
3	tweet similarity	187	1578	372	1763	0.335	0.096	0.149	-0.135
4	URL Ratio	967	1920	26	893	0.974	0.520	0.678	0.577
5	friends/(followers)^2	1816	1850	95	90	0.950	0.953	0.952	0.904
Yang et al. [25]									
1	age	308	1420	746	1455	0.292	0.175	0.219	-0.191
2	bidirectional link ratio	1888	1927	18	24	0.991	0.987	0.989	0.978
3	average neighbors' followers	1909	648	1299	38	0.595	0.980	0.741	0.411
4	average neighbors' tweets	1291	1518	430	616	0.750	0.677	0.712	0.459
5	followings/median neighbor's followers	1926	916	1031	21	0.651	0.989	0.785	0.538
6	API ratio	1431	1917	29	429	0.980	0.769	0.862	0.776
7	API URL ratio	141	1851	66	288	0.681	0.329	0.443	0.401
8	API tweet similarity	87	2157	9	1676	0.906	0.049	0.094	0.146
9	following rate	1597	1458	492	353	0.764	0.819	0.791	0.568
Camisani–Calzolari [4] (rules detecting human behavior)									
1	profile has name	0	1950	0	1950	—	—	—	—
2	profile has image	2	1931	19	1948	0.095	0.001	0.002	-0.06
3	profile has address	323	1313	637	1627	0.336	0.166	0.222	-0.187
4	profile has biography	617	1806	144	1333	0.811	0.316	0.455	0.306
5	followers $\geq$ 30	1852	1582	368	98	0.834	0.95	0.888	0.768
6	belongs to a list	1893	1052	898	57	0.678	0.971	0.799	0.566
7	tweets $\geq$ 50	1582	1792	158	368	0.909	0.811	0.857	0.735
8	geo-localization	1923	678	1272	27	0.602	0.986	0.748	0.434
9	has URL in profile	1895	697	1253	55	0.602	0.972	0.743	0.417
10	in favorites	1130	1748	202	820	0.848	0.579	0.689	0.502
11	uses punctuation in tweets	93	1948	2	1857	0.979	0.048	0.091	0.151
12	uses hashtags	437	1934	16	1513	0.965	0.224	0.364	0.337
13	uses iPhone to log in	1905	845	1105	45	0.633	0.977	0.768	0.489
14	uses Android to log in	1932	677	1273	18	0.603	0.991	0.75	0.442
15	has connected with Foursquare	1943	261	1689	7	0.535	0.996	0.696	0.257
16	has connected with Instagram	1890	772	1178	60	0.616	0.969	0.753	0.446
17	uses the website Twitter.com	131	1852	98	1819	0.572	0.067	0.12	0.036
18	has tweeted a userID	1071	1941	9	879	0.992	0.549	0.707	0.609
19	2*followers $\geq$ friends	1947	863	1087	3	0.642	0.998	0.781	0.531
20	tweets do not just contain URLs	125	1943	7	1825	0.947	0.064	0.12	0.167
21	retwitted tweets $\geq$ 1	1021	1915	35	929	0.967	0.524	0.679	0.569
22	uses different clients to log in	118	1924	26	1832	0.819	0.061	0.113	0.125
Van Den Beld (State of search) [17]									
1	*bot* in biography	0	1950	0	1950	—	—	—	—
2	following:followers = 100:1	158	1950	0	1792	0.541	1	0.15	0.205
3	same sentence to many accounts	188	1521	429	1762	0.438	0.78	0.146	-0.169
4	duplicate profile pictures	26	1809	141	1924	0.471	0.928	0.025	-0.146
5	tweet from API	429	33	1917	1521	0.118	0.017	0.2	-0.779
Socialbakers [16]									
1	friends:followers $\geq$ 50:1	316	1949	1	1634	0.997	0.162	0.279	0.296
2	tweets spam phrases	5	1950	0	1945	1	0.003	0.005	0.036
3	same tweet $\geq$ 3	30	1327	623	1920	0.046	0.015	0.023	-0.407
4	retweets $\geq$ 90%	14	1933	17	1936	0.452	0.007	0.014	-0.009
5	tweet-links$\geq$ 90%	58	1936	14	1892	0.806	0.03	0.057	0.084
6	0 tweets	84	1949	1	1866	0.988	0.043	0.083	0.146
7	default image after 2 months	2	1931	19	1948	0.095	0.001	0.002	-0.06
8	no bio, no location, friends $\geq$100	255	1927	23	1695	0.917	0.131	0.229	0.231

Table 1: Rules evaluation

candidate rules for training classifiers for fake follower detection.

5.2.1 Camisani-Calzolari detection algorithm

Table 2 reports the results of the detection algorithm proposed in [4], aggregating the 22 criteria for human/bot behavior detection. In particular, the algorithm evaluates every single rule on each account and assigns a positive human score and a negative bot score, based on the single rule output, as described in Appendix A.3. The final outcome depends on the score obtained by the account: if the result is a score greater than 0, then the

dataset	real humans	outcome		
		humans	bots	neutral
@TheFakeProject	469	456	3	10
#elezioni2013	1481	1480	0	1
100% fake	0	2889	185	277

Table 2: Camisani-Calzolari algorithm outcomes on our gold standard datasets

account is marked as "human"; if it is between 0 and -4 it is marked as "neutral", otherwise it is marked as "bot".

After applying the algorithm to our mixed dataset, we observe that, although obtaining very good results in detecting the real human accounts, it achieves a poor fake account detection. Most of them have been erroneously tagged as humans too. The main motivation of this unsatisfactory result is that our fake follower accounts feature characteristics that easily make them obtaining a human score higher than the bot one. This strengthens the thesis of this paper: detecting fake followers is a specific research topic deserving further investigations.

6. TAXONOMY OF FOLLOWERS

In this section, we present a classification of Twitter followers based on multiple dimensions derived from results exposed in previous sections.

Categories. We start considering the following three categories for Twitter followers:

1. *Humans*
2. *Bots*, i.e., automated programs, that can be further partitioned in:
   - *benign*: those automated accounts programmed to, e.g., regularly post a large volume of benign tweets, like news and blogs updates [5]
   - *malicious*:, those automated programs mainly exploited to spread spam and malicious contents. In the following, similar to [21], we refer to this sub-category as *spambots*.
3. *Fake followers*, either automated programs, or human accounts. Note that an human account can be classified as fake when the owner of the account has been victim of an attack finalized at stealing her authentication credentials, as highlighted in [20], and these credentials have been later used by a third party.

Dimensions. We propose to consider the following six dimensions, with the aim to characterize followers as belonging to one of the above introduced categories.

1. *Automation*, indicating the level of automation of the account (i.e., fully, partly, or not automated).
2. *Purpose of use*, that is, the motivations that drive a user to use Twitter.
3. *Link rationale*, that is, the rationales motivating why account a chooses to follow account b.
4. *Tweet number,* i.e., a measure of how many tweets an account posts over a period of time.
5. *Audience*, indicating the modality employed by an account to post its messages (either tweets, hence broadcast to all its followers, or direct messages to another account, not publicly visible).
6. *Expression of interest*, denoting how the account relates and interacts with other ones. This dimension can be measured by, e.g., the degree of occurrence of mentions and replies in those tweets posted by the account, or by the number of times an account's tweet is a retweet.

Below, we discuss the three categories through the above-introduced dimensions. We also show a qualitative and visual comparison through three radar graphs reported in Figure 1.

Automation. A classification of Twitter followers based on their automation level has been given in [5], where the authors differentiate among bots, humans, and cyborgs. While bots are fully automated programs, humans feature manual behavior, and cyborgs interweave characteristics of both manual and automated behavior. For example, a human may register an account and set automated programs to post tweets during her absence. According to [20], that explains the dynamics around Twitter markets, and from the analysis of our datasets of fakes, we also argue that fake followers can be either automated programs, or human accounts (with a preponderance of the former). Moreover, it is worth noticing how the human fake followers usually follow the target account without being aware of it, either because they have given away (on a voluntary basis) their credentials (that are later used by a third party) or because they are victims of Twitter Account Markets [20]. Given these considerations, we decided to associate a very high level of automation to spambots, a very low one to human accounts, and an average one to fake followers, see Figure 1.

Purpose of use. Works in [20, 21, 25] on spam detection identify, as typical activities for spambots on Twitter, the sending of unsolicited promotional tweets to, e.g., advertise questionable websites or products; the spread of malware; soliciting illicit activities [13] On the contrary, humans' main use of Twitter is for microblogging, quick answers from crowd, keeping up with the news, finding out what other people think about politics, books, a particular brand of body wash, etc. [1].

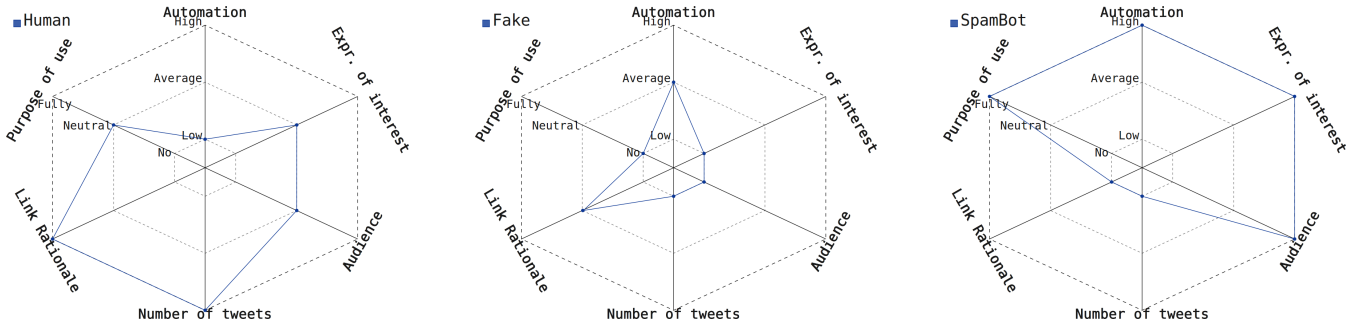

Figure 1: Human, Fake and Spambot classification through six dimensions.

Finally, fake followers are those followers explicitly created with the goal of *inflating the number of follower of a target account*, contributing in such a way to increase popularity and online reputation of the target, see, e.g. [20]. To graphically render this dimension, we decided to map it to the *bad behavior* axis in the radar-fashion Figure 1, thus achieving a fully bad behavior for spambots, a very low level (even null) for humans, and a neutral level for fake followers.

Link rationale. By construction, spambots do not follow a particular rationale to follow some particular accounts, in contrast with human accounts. This is discussed more formally in [25], that model the Twittersphere as a graph where each account is a node, and every following/follower relationship is an edge. Work in [25] notes that, "since spammers usually blindly follow other accounts, these accounts usually do not know each other and have a looser relationship among them", thus, they do not form a clique; in addition, following such unrelated accounts, the spambot will create new shortest paths between those that re-follow it, leading to a position in the graph more central for the spambot than for human accounts. Finally, the number of bidirectional links has been shown to be higher for an human account than for a spambot (see Appendix A.2 for details). As it can be derived from our analysis – Section 5 – some conclusions that are valid for spambots also apply to fake followers: bidirectional links help to recognize fake followers since usually the followed account has no interest in following back a fake. This leads us to conclude that graph-based features can help in understanding the rationale of fakes. When comparing spambots, fake and humans, we can say that spambots and fake are more prone to randomly follow other accounts, whereas humans are more prone to form cliques, since friends and relatives usually follow each other. Finally, the link rationale that causes fake followers to follow who paid for this service (being followed) could be the only element to distinguish the random link rationale of spambots. Graphically, we assign a high link rationale to humans, low link rationale to spambots, and an average link rational to fakes, see Figure 1.

Tweet number. In [21], the authors created a number of honey profiles on Twitter, that received, over an 11 months period, about 360 friend requests from spambots. By analyzing the activity of the spambots, the authors conclude that, compared to legitimate users (humans), they are less likely to send hundreds of tweets (most of them sent less than twenty messages). To support this observation, note that also the authors of [25] noticed that all spambots under investigation sent a much lower number of tweets than legitimate accounts. However, spambots can use specialized softwares to automatically post tweets. From the analysis of our results, the number of tweets of the fakes is also significantly less than that of humans. Indeed, only 19% of the fakes in our mixed dataset have written at least 50 tweets (at the time those accounts have been crawled), compared to an average of 88% of the human accounts in #elezioni2013 (95%) and @FakeProject (81%). We conclude that both fakes and spambots usually send a quite lower number of tweets with respect to human accounts.

Audience. In [21], the authors find that 20 out of the 360 spambots following their honey profiles have sent direct messages (DMs) to those profiles. Such spambots have been denoted as *whisperers*, in contrast to *braggers*, this latter ones being spambots that spam through public tweets. Surfing the Web, one can easily find tools featuring the cleaning up of Direct Messages Inbox from spam, e.g., [10]. No tool is provided by Twitter to monitor sent and received DMs. Thus, an exact evaluation of this dimension is highly challenging. However, our sets of 3000 fakes follow three distinct honey accounts, and none of them has ever received a direct message from any of the following accounts. Thus, we may conclude that a difference exists between a spambot and a fake follower behavior. In particular, in their attitude to make use of direct messages. Reasonably, we can also assert that the account attitude to send direct messages could be definitely exploited by Twitter to spot spam-like (and human-like) behaviours. For classification purposes, we conclude that spambots send more

DMs than humans, that, in turn, send more DMs than fakes. However, this analysis is preliminary and deserves further investigation.

Expression of Interest. We may define several metrics to weigh this dimension, e.g., the *mention* ratio, i.e., $\frac{tweets\ with\ mentions}{total\ tweets}$ (the number of tweets containing mentions over the total number of tweets for an account); the *retweet* ratio, i.e., $\frac{retweeted\ tweets}{total\ tweets}$ (the number of tweets that are retweets of other users tweets over the total number of tweets for an account); and the *reply* ratio, i.e., $\frac{replies}{total\ tweets}$ (the number of tweets that are replies over the total number of tweets for an account). Interestingly, two of these three metrics have been used to detect spambots in [24, 9], where quantitative results on training datasets show that the values for these ratios are significantly higher for spambots than for humans. We have adopted the rule of information gain with continuous attributes to test such results on our mixed dataset, and we have achieved an opposite result: none of our fakes have ever replied to some other tweets (threshold = 0). In addition, the mention ratio threshold is quite low, 0.18 (if *mention* ratio $\leq$ 0.18, the account in our mixed dataset is a fake follower). We have also computed the threshold for the *retweet* ratio (if *retweet* ratio $\leq$ 0.05, the account in our mixed dataset is classified as a fake follower), but we cannot compare this last result with a measure for spambots, since we were unable to find related work in the literature considering this metric. We can thus conclude that the expression of interest is higher for spambots, lower for fakes, while human accounts lie in between.

The list of suggested dimensions is not exhaustive. As an example, one could adopt as a dimension the way tweets are stylistically composed (e.g., if all tweets by an account are written in the same language). Although other dimensions could have been caught as well, we believe that the six ones reported above help to capture the main features needed to characterize an account.

Finally, while conclusions drawn around some dimensions are highly supported by the current analysis and past investigations, some others, while being plausible deductions, call for further research. Table 6 classifies the six dimensions based on how much our proposed taxonomy is supported by experimental data.

dimension	support
Automation	high
Purpose of Use	high
Link Rationale	medium
Number of Tweets	medium
Audience	low
Expression of Interest	medium

Table 3: Support per dimension

7. CONCLUSION AND FUTURE WORK

In this paper, we have shed light on the lack of rigorous definitions and criteria for the identification of fake account on Twitter. Further, we have also provided the basis for further investigation in this direction.
In particular, to reach the first goal, we created two gold standard; one of human and one of fake accounts. Aiming at detecting the latter ones, given the lack of specific criteria, we identified a set of criteria proposed in the literature as well as in the media to target spambots and inactive accounts, and we tested their effectiveness on our reference dataset. Noticeably, just few criteria succeeded in correctly detecting the fake followers. This result supported our thesis that fake detection represents an open research issue.
A promising starting point to develop automatic classifiers for fake detection is provided by the criteria identified in this paper. A further contribution was to create a novel taxonomy, comparing humans, spambots, and fakes, through multiple dimensions. The taxonomy gives us hints on which dimensions are more relevant to discriminate between fakes and spambots.
As future work, we aim to design an automated fake detector engine, leveraging the dimensions proposed in our taxonomy and exploiting the main differences we found between the different account categories.

8. REFERENCES

[1] About.com. 10 Great Uses For Twitter. In *http://goo.gl/q4bjb*. Last checked June 27, 2013.

[2] P. Baldi, S. Brunak, Y. Chauvin, C. Andersen, and H. Nielsen. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424, 2000.

[3] Brandwatch.com. Analysis of global brands' Twitter activity. In *http://goo.gl/C6MeU*, Dec. 2012. Last checked June 27, 2013.

[4] M. Camisani-Calzolari. Analysis of Twitter followers of the US Presidential Election candidates: Barack Obama and Mitt Romney. In *http://digitalevaluations.com/*, August 2012.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. Dependable Sec. Comput.*, 9(6):811–824, 2012.

[6] Corriere Della Sera (online ed.). Academic Claims 54% of Grillo's Twitter Followers are Bogus. In *http://goo.gl/qi7Hq*, July 2012. Last checked June 27, 2013.

[7] Financial Times – Tech blog (online ed.). Twitter bots are boosting brands - survey. In *http://goo.gl/Zt2l2*, June 2012. Last checked June 27, 2013.

[8] La Repubblica (online ed.). Twitter, quanti falsi profili: il CNR ora va a caccia dei fake. In

http://goo.gl/zNC2k, December 2012. Last checked June 27, 2013.

[9] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In *SIGIR Conference on Research and Development in Information Retrieval*, pages 435–442. ACM, 2010.

[10] S. Malarkey. How to get rid of DM spam on Twitter. In *http://goo.gl/P259*. Last checked June 27, 2013.

[11] T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.

[12] New York Times (online ed.). Buying Their Way to Twitter Fame. In *http://goo.gl/VLrVK*, August 2012. Last checked June 27, 2013.

[13] Official Twitter Blog. Avoid Phishing Scams. In *http://goo.gl/zOgHi*, February 2010. Last checked June 27, 2013.

[14] Skande.com. Twitter: un progetto del CNR cerca i Fake Followers #ImNotAFake. In *http://goo.gl/V3zX6*, December 2012. Last checked June 27, 2013.

[15] C. Smith. By The Numbers: 16 Amazing Twitter Stats. In *http://goo.gl/2Xr9X*, May 2013. Last checked June 27, 2013.

[16] SocialBakers. Fake follower check. In *http://goo.gl/chWn0/*. Last checked June 27, 2013.

[17] Stateofsearch.com. How to recognize Twitterbots: 7 signals to look out for. In *http://goo.gl/YZbVf*, September 2012. Last checked June 27, 2013.

[18] Statistic Brain. Twitter statistics. In *http://goo.gl/XEXB1*, May 2013. Last checked June 27, 2013.

[19] Statuspeople.com. Status People Fakers. In *http://goo.gl/0Jpky*. Last checked June 27, 2013.

[20] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna. Poultry markets: on the underground economy of Twitter followers. In *Workshop on online social networks*, WOSN '12, pages 1–6. ACM, 2012.

[21] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9. ACM, 2010.

[22] The Guardian (online ed.). Barack Obama tweets the start to his 2012 re-election campaign. In *http://goo.gl/Uk6Av*, April 2011. Last checked June 27, 2013.

[23] The Telegraph (online ed.). Human or 'bot'? Doubts over Italian comic Beppe Grillo's Twitter followers. In *http://goo.gl/2yEgT*, July 2012. Last checked June 27, 2013.

[24] A. Wang. Don't follow me: Spam detection in Twitter. In *Intl. Conference on Security and Cryptography (SECRYPT)*, pages 1–10, 2010.

[25] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers. In *14th International Conference on Recent Advances in Intrusion Detection*, RAID'11, pages 318–337. Springer-Verlag, 2011.

[26] S. Yardi, D. M. Romero, G. Schoenebeck, and D. Boyd. Detecting Spam in a Twitter Network. *First Monday*, 15(1), 2010.

APPENDIX

A. CRITERIA DETAILS

Here, we detail those criteria proposed in [21, 25, 4, 17, 16] for spam/bot detection on Twitter. As described in Section 5, such criteria have been applied to our gold standard to investigate their strength in discerning humans and fake followers.

For each group of criteria we also detail how we address, at implementation level, some ambiguities and/or technical problems that would have, potentially, affected the final outcome of the evaluation. For each of some issues, we provide a note explaining how it was addressed.

Hereafter, we report the thresholds used for all those rules in Table 4 requiring a comparison of the measured value with a fixed value. Such thresholds have been fixed according to the information gain rule [11].

	criterion	account is fake when
Stringhini et al. [21]		
4	URL Ratio	≤ 0.056051
5	friends/(followers)^2	> 0.493827
Yang et al. [25]		
2	bidirectional link ratio	≤ 0.064541
3	average friends' followers	≤ 389913
4	average friends' tweets	≤ 3068.6
5	friends/median friends' followers	> 0.078602
6	API ratio	≤ 0
7	API URL ratio	> 0.997992
9	following rate	> 0.631206

Table 4: Thresholds used to evaluate the rules

A.1 Detecting spammers in social networks

The authors in [21] identify the following characteristics, that allow a classifier to detect spambots on Twitter. The account under investigation may show a spambot behavior if:

1. it does not have thousands of friends;
2. it has sent less than 20 tweets;
3. the content of its tweets feature the so called *message similarity*;
4. it has a high $\frac{tweets\ containing\ URLs}{total\ tweets}$ ratio (*URL ratio*);

5. it has a high $\frac{friends}{(followers)^{\wedge}2}$ ratio value (i.e., lower ratio values mean legitimate users).

Notes.

For the rule 1 we fix the threshold equal to 500.

For the rule 3 we implemented the notion of message similarity by checking the existence of at least two tweets, in the last 15 tweets of the account timeline, in which 4 consecutive words are equal (words are consecutive characters separated by white spaces). This notion has been given in a latter work by the same authors, see [20].

For the rules 4 and 5 we used the thresholds in Table 4, obtained as detailed in Section 5.

A.2 Fighting evolving Twitter spammers

The authors of [25] propose a set of metrics that should be highly robust against spammer evasion techniques. Hereafter, we detail nine of their properties:

1. *age*: the more an account is aged, the more it could be considered a good one (this rule also appears in [9]);
2. *bidirectional link ratio* is $\frac{bidirectional\ links}{friends}$, where a bidirectional link is when two accounts follow each other among them. This ratio has been tested to be lower for spammers accounts than for legitimate accounts;
3. *average neighbors' followers* represents the average number of followers of the account's friends, and it aims at reflecting the quality of the choice of friends of an account. It is commonly higher for legitimate accounts than for spammers;
4. *average neighbors' tweets*: the average number of tweets of the account's followers should be lower for spammers than for legitimate accounts;
5. *followings to median neighbor's followers* of an account is defined as the ratio between the number of friends and the median of the followers of his friends: this value has been found higher for spammers than for legitimate accounts;
6. *API ratio* (= n.tweets sent from API / total n. tweets): work in [5, 25] reveal higher values for suspicious accounts;
7. *API url ratio* (= n.tweets posted from API and containing URL / total n.tweets posted from API). This value is higher for suspicious accounts;
8. *API tweet similarity*: this metric considers the number of similar tweets sent from API. The notion of tweet similarity is as in Section A.1;
9. *following rate*: this metric reflects the speed at which an accounts follows other accounts, and the idea is that higher values are related to spammers.

Interestingly, the authors of [25] also suggest two further graph-based metrics. Seeing the Twittersphere as a graph, where accounts are nodes, and each follow relationship is an edge, the first metric is the *local clustering coefficient*, that quantifies how close the neighbors of a node in a graph are to be a clique. The intuitive idea behind this metric is that spammers blindly follow other accounts, that do not know each other and have a looser relationship among them, thus, they do not form a clique. Thus, spammers have lower local clustering coefficients, compared to humans. The second metric is the *betweenness centrality* which reflects the position of a node in the graph (namely how much a node is involved in the shortest paths between all the possible vertices pair). The intuitive idea behind this metric is that, following unrelated accounts, the spammer will create new shortest paths between those who re-follow it, leading to a position in the graph more central for the spammer than for human accounts.

Although these metrics should be very effective to recognize spammers, unfortunately they are extremely expensive to evaluate and the same authors evaluated it using a sampling techniques. This is the reason why we have not implemented them in our analysis, and, thus, how these metrics behave in discriminating fake followers is an open issue.

Finally, the authors review other metrics claimed to be less robust against evasion techniques. For this reason, we decided not to include them in our evaluation.

Notes.

Precisely evaluating rule 9 requires to know the evolution of the number of friends of an account. Actually this kind of information is publicly unavailable. Thus, as in [25], we also approximate the rate as friends / age.

A.3 Followers of political candidates

The technical report [4] aims at detecting bot or inactive accounts, using a series of tests. The study assigns to the examined accounts 1 (or more, where specified) human (or "active") point for each of the following criteria:

1. the profile contains a name;
2. the profile contains an image;
3. the profile contains a physical address;
4. the profile contains a biography;
5. the account has at least 30 followers;
6. it has been inserted in a list by other Twitter users;
7. it has written at least 50 tweets;
8. the account has been geo-localized;
9. the profile contains a URL;
10. it has been included in another user's favorites;
11. it writes tweets that have punctuation;

12. it has used a hashtag in at least one tweet;
13. it has logged into Twitter using an iPhone;
14. it has logged into Twitter using an device with Android ;
15. it has connected with Foursquare;
16. it has connected with Instagram;
17. it has logged into *twitter.com* website;
18. it has written the userID of another user in at least one tweet;
19. 2*number followers $\geq$ number of friends;
20. it publishes content which does not just contain URLs;
21. at least one of its tweets has been re-twitted by other accounts (it's worth 2 points);
22. it has logged into Twitter through different clients (it's worth 3 points).

Moreover, for each criterion not verified, the account receives 1 bot (or "inactive") point, with the exception of criteria 13, 14, 15, 16 and 17. For the criterion 21, 2 bot points are assigned if no tweets of the account have been retweeted by other users.

Notes.

For the rule 8, *geo-localization* is related to tweets. Thus, we consider this rule satisfied when at least one tweet of the account has been geo-localized.

For the rule 11, *punctuation* has been searched in both the profile biography and in its timeline.

A.4 Blogs

The "7 signals to look out for" recognizing Twitter bots are the following [17]:

1. the biography of the profile clearly specifies that it is a bot account;
2. the friends to followers ratio is in the order of 100:1;
3. the account tweets the same sentence to many other accounts;
4. different accounts with duplicate profile pictures are suspicious;
5. accounts that tweet from API are suspicious;
6. the response time (follow+reply) to tweets of other accounts is within milliseconds;
7. the account tends to follow/unfollow other accounts within a temporal arc of 24 hours.

Notes.

We did not apply rules 6 and 7 to our datasets, since they require to actively interact with the account. This means that those rules cannot be used for an automatic process of fake detection.

The rule 3 has been implemented considering the tweet as a single unit (contrary to tweet similarity, that focuses on the occurrence of the same four consecutive words in two or more tweets). The last 20 tweets of each timeline have been considered.

For the rule 4, we consider a duplicate profile picture when at least 3 accounts within a dataset have the same profile picture.

For the rule 5, we consider as tweets posted from API all those tweets not being posted from the website *twitter.com*.

A.5 Socialbakers Fake Follower Check

The Fake Follower Check tool by the social media firm Socialbakers [16] evaluates the fakeness of an account followers using the following criteria:

1. $\frac{friends}{followers}$ ratio is 50:1 or more;
2. more than 30% of all tweets use spam phrases, such as "diet", "make money" and "work from home";
3. the same tweets are repeated more than three times, even when posted to different accounts;
4. more than 90% of the account tweets are retweets;
5. more than 90% of the account tweets are links;
6. the account has never tweeted;
7. the account is more than two months old and still has a default profile image;
8. the user didn't fill in neither bio nor location and, at the same time, is following more than 100 accounts.

The Socialbakers website reports the Fake Follower Check as a beta version, adding the following: "We are currently tweaking the algorithm". Thus, we consider the criteria published on the firm website at time of writing, in June 2013.

Notes.

For the rule 2, we consider as spam phrases expressions like "diet" or "make money" or "work from home" (both English and Italian translations), as suggested by the website of Socialbakers.