Spotting Fake Retweeting Activity in Twitter

Maria Giatsoglou*, Despoina Chatzakou*,Neil Shah†,Alex Beutel†,Christos Faloutsos†,Athena Vakali*
*Informatics Department, Aristotle University of Thessaloniki, Greece
{mgiatsog, deppych, avakali}@csd.auth.gr
†School of Computer Science, Carnegie Mellon University, USA
{neilshah, alexbeutel, christos}@cs.cmu.edu

***Abstract*—Given the retweeting activity around the posts of several Twitter users, how can we spot organic users' reactions from fake retweets that aim to boost a post's appearance of popularity? Our main intuition is that organic behavior has more variability, while fraudulent behavior is more *synchronized*. We refer to the detection of such fraudulent activities as the *Retweet Fraud* problem and propose: (A) a set of carefully designed features for characterizing retweet threads, i.e. a given post's retweets, (B) an efficient method for spotting fraudsters. We present experiments on a real dataset of almost 12 million retweets crawled from Twitter, where our method achieves a 97% accuracy.**

I. Introduction

The abundance of content in Twitter is of varying quality - however, interesting and attractive information is typically promoted via the *retweet* action for re-broadcasting existing tweets. High retweet activity is considered as an indication of the trustworthiness of information and the influence of its publisher, and it can be monetized by per-page or per-click advertisements. Thus, Twitter has become a host to many fraudulent users that aim to falsely create the impression of popularity by artificially generating a high volume of retweets for their posts. Such users can be of different types depending on their: motivation (product/service/user promotion, malware/spam), account's automation-level (automated or bot orchestrated, semi-automated, human managed) and fraud practice (self-owned/hired bot network, retweets as-a-service). Fraudulent users' activities blend with those of legitimate users and undermine the credibility of content in Twitter.

Thus, it is important to find features that can separate such diverse fraudulent activities from honest user behavior, and design a method that can effectively use them for revealing the various types of fraud. Our main intuition is that, unlike organic behavior, suspicious activities are highly *synchronized*, i.e. they exhibit the patterns across time. Synchronized ("lock-step") behavior has been often used as an indication of fraud, like e.g. users Liking the same Facebook Pages at the same time [1] or following the same accounts [3]–[5].

Here, we propose an effective method that leverages synchronicity for addressing the RETWEET FRAUD problem, i.e. the detection of Twitter users that obtain fake retweets for their posted content (i.e. RT fraudsters). We anticipate that such RT fraudsters will be exposed by the high synchronicity of their retweet threads—a *retweet thread* is the set of tweets that have retweeted a given post. E.g., consider a bot network where each bot reweets with 1 second from the previous, or a user who regularly "buys" 1,000 retweets from online retweet markets. We define the RETWEET FRAUD problem as follows.

*Problem 1 (*RETWEET FRAUD*):*

Given: a set of Twitter users and a set of retweet threads for each user, represented in a p-dimensional feature space
Extract: a set of features at the user-level, and
Identify: RT fraudsters, i.e. users exhibiting abnomaly *synchronized* characteristics w.r.t. their retweet threads.

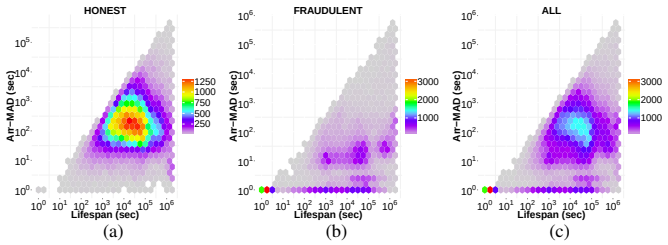

Fig. 1: **Synchronicity patterns are revealed as microclusters in RT fraudsters behavior**. a, b, c correspond to the Lifespan vs. Arr-MAD scatter plots for "honest", "fraudulent", and both types of users.

The contributions of this work are the following:

- **Feature engineering**: we propose a carefully selected set of features for retweet threads, that are suitable for spotting synchronized retweeting behavior;
- **Methodology**: we propose an effective and generalizable method that automatically detects the different manifestations of RETWEET FRAUD.

II. Proposed approach

The proposed approach entails a 7-dimensional feature space for retweet threads and a data analysis pipeline.

A. Descriptive Features for Retweet Threads.

Our goal was to identify features whose values are abnormally coherent over the retweet threads of a given fraudulent user, compared to the expected, normal behavior. After experimenting with several features, we ended up with a set of features based on the timing and quantity of retweets within a retweet thread. The suitability of such features for the RETWEET FRAUD problem is supported by the typical use of bot-like automation tools for (re)tweeting, and due to the way retweet markets operate (e.g. bulk purchase of a fixed number of retweets). Table I summarizes the proposed features.

TABLE I: Proposed feature set

Feature	Description
Retweets	number of retweets
Response time	time elapsed between the tweet's posting time and its first retweet
Lifespan	time elapsed between the first and last (observed) retweet, constrained to 3 weeks to remove bias
RT-Q3 response time	time elapsed after the tweet's posting time to generate the first 3 quarters of the (observed) retweets
RT-Q2 response time	time elapsed after the tweet's posting time to generate the first half of the (observed) retweets
Arr-MAD	mean absolute deviation of retweets' inter-arrival times
Arr-IQR	inter-quantile range of retweets' inter-arrival times

We examined the proposed features' suitability as indicators of RETWEET FRAUD by projecting several retweet threads (the dataset is described in Section III) in all 2-D feature subspaces derived by the proposed feature set. Figure 1, indicatively, depicts the binned 2-D heatmap, in logarithmic scales, of *Lifespan vs. Arr-MAD* for all (Fig. 1c), honest (Fig. 1a), and

fraudulent (Fig. 1b) users. We can easily discern microclusters of fraudulent retweet threads (at very low values of Arr-MAD and at high Lifespan values), whereas the majority of honest users' retweet threads are concentrated around a certain area of the feature subspace, clearly separated from fraudsters. Similar observations were made on the other heatmaps.

B. Method Pipeline.

Our method comprises three steps: (1) *Feature subspace sweeping*; (2) *User scoring*; (3) *Multivariate outlier detection*.
Feature subspace sweeping. Given N users with their set of retweet threads represented in a $p-$feature space, we first project all threads into all 2^p possible feature subspaces, and then segment each subspace by applying logarithmic binning, in powers of 2, in each dimension. The projection of retweet threads in all possible feature subspaces allows us to address diverse types of fraud, characterized by the appearance of micro-clusters for different combinations of features.
User scoring. We combine the thread-level features into user-level scores reflecting the threads' synchronicity. For each user, we calculate a *suspiciousness score* on each retweet thread-feature subspace f_i, $i \in [1, 2^p]$. This score indicates how coherent a given user's retweet threads are, when projected in f_i. Coherence is computed as the average closeness between the projections of all pairs of a user's threads, where closeness is a binary function: it is 1 when the threads belong to the same bin, and zero otherwise. Then, the suspiciousness score is taken as the residual of coherence from a lower bound. Finally, each user is represented by a *suspiciousness vector* compising the scores over all feature subspaces.
Multivariate outlier detection. Then, the method proceeds to the identification of RT fraudsters based on the users' suspiciousness vectors. We spot as suspicious the users that largely deviate from the majority, considering their standardized scores in all feature subspaces, based on a robust approach for multivariate outlier detection [2]. This approach finds a robust feature subspace V_r that fits the majority of users, and detects as outliers users whose orthogonal distance from V_r (od score), or the distance of their projection from the majority of users on V_r (sd score), surpasses a data-adaptive cutoff.

III. Experiments and Results

We test our method on a dataset crawled from Twitter which comprises over *130K retweet threads* summarizing $\sim$ *12M retweets* to posts of 298 active Twitter users. Due to the Twitter API's constraint of partial access to the published tweets, our requirement for *complete* retweet threads (no gaps in a post's retweets), and the lack of a relevant (labeled) dataset, we manually selected and annotated a set of target users (90% "'honest"; 10% "fraudulent") and tracked all tweets and their retweets over a time period. Target users were selected in several fashions: (A) we studied a 2-day sample of the global Twitter timeline and sampled users who posted: the most retweeted tweets and tweets containing keywords heavily used in spam campaigns (*casino*, *followback*, etc); (B) based on a web site that publishes rankings of Twitter users in terms of popularity, we randomly chose users who tweeted several times per week and had $>$ 100 retweets on their recent posts; (C) we collected users active in specific topics (European affairs and Automobile), given that they were added in such topic-related lists by other users. Then we manually labeled users as "fraudulent" if (i) inspection of their tweets' content revealed spammy links to external web pages, spam-related keywords, and multiple posts with similar promotions

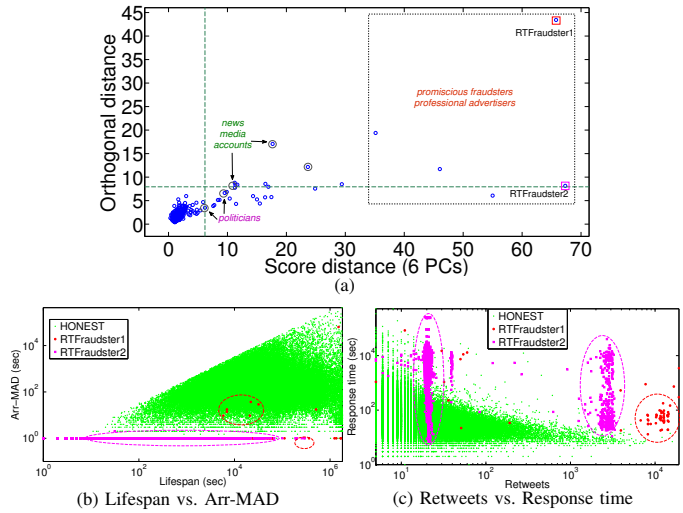

Fig. 2: **Our method detects outliers clearly separated from the majority of users.** In (2a), points right of the vertical dashed line are spotted as RT fraudsters. (2b), (2c) reveal high synchronicity in the retweet threads of users RTFraudster1 and RTFraudster2, for different feature combinations.

or vacuous content (e.g. quotes); (ii) profile information was clearly fabricated/indicated promotion activity.

The proposed method was very effective on the detection of the dataset's RT fraudsters, achieving **97% accuracy** and **0.82 F1-score**. Figure 2a illustrates the *outlier map*, where each point is a user plotted in terms of the derived sd and od scores. The red dashed lines correspond to their adaptive cutoff values – the plot clearly discerns the outliers from the majority of users which lie in the bottom-left region.

A closer examination of the caught RT fraudsters reveals that the ones who scored high in sd and od (at the rightmost part of the figure) were exemplary bot accounts, typically hired for promotion or advertisement. E.g., figures 2b and 2c depict the retweet threads of normal users (green points) with those of the "significantly outlying" caught users RTFraudster1 and RTFraudster2 projected in 2-D subspaces with respect to two pairs of our proposed features. Compared to honest retweet threads, we can observe that fraudulent users' retweet threads are abnormaly clustered together. RTFraudster1 is in fact a *promiscuous* fraudster with 800 followers that had 65 retweet threads in a 4-month time period – 80% (60%) of these comprised more than 1k (10k) retweets and had $\tilde{0}$ Arr-IQR.

The remainder of the "caught" RT fraudsters have a more *subtle* profile, resembling cyborg behavior: they often create vacuous posts, but occasionaly interact with other users, indicating a human operator. We found that the 5 false positives detected by our method (enclosed by a red circle in Figure 2a) belong to media accounts and politicians. Three of these accounts have significantly abnormal sd and od scores, while the others are situated very close to RT fraudsters, suggesting that they may have tampered with their retweet threads.

References

[1] A. Beutel, et al. CopyCatch: stopping group attacks by spotting lockstep behavior in social networks. In *WWW*, 119–130. ACM, 2013.
[2] M. Hubert, et al. Robust PCA for Skewed Data and its Outlier Map. *Computational Statistics & Data Analysis*, 53(6):2264–2274, 2009.
[3] M. Jiang, et al. CatchSync: Catching Synchronized Behavior in Large Directed Graphs. In *KDD*, 941–950. ACM, 2014.
[4] M. Jiang, et al. Inferring Strange Behavior from Connectivity Pattern in Social Networks. In *PAKDD*, Tainan, Taiwan, 2014.
[5] N. Shah, et al. Spotting Suspicious Link Behavior with fBox: An Adversarial Perspective. In *ICDM*, 2014.