Retweeting activity on Twitter: Signs of Deception

Abstract. Given the re-broadcasts (i.e. retweets) of posts in Twitter, how can we spot fake from genuine user reactions? What will be the tell-tale sign — the connectivity of retweeters, their relative timing, or something else? High retweet activity indicates influential users, and can be monetized. Hence, there are strong incentives for fraudulent users to artificially boost their retweets' volume. Here, we explore the identification of fraudulent and genuine retweet threads. Our main contributions are: (a) the discovery of *patterns* that fraudulent activity seems to follow (the "TRIANGLES" and "HOMOGENEITY" patterns, the formation of micro-clusters in appropriate feature spaces); and (b) "RTGEN", a realistic generator that mimics the behaviors of both honest and fraudulent users. We present experiments on a dataset of more than 6 million retweets crawled from Twitter.

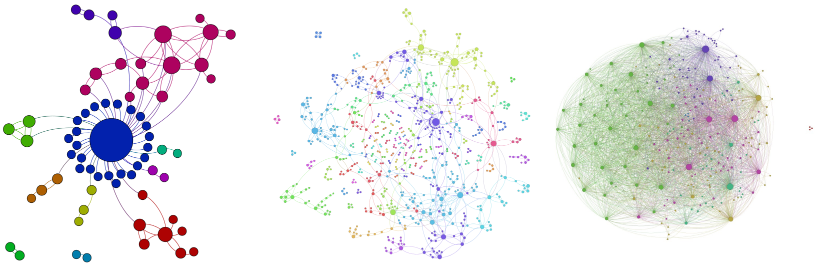
1 Introduction

Can we spot patterns in fake retweeting behavior? When a large number of Twitter users re-broadcast a given post, should we attribute this burst of activity to organic, genuine expression of interest or rather to a fraudulent, paid contract? Twitter is arguably the most popular micro-blogging site and one of the first sites forbidden by authoritarian regimes. High-quality tweets are re-broadcasted (*retweet*ed) by many users, indicating that their authors are influential. Since such influence can be monetized via per-click advertisements, Twitter hosts many fraudsters trying to falsely create the impression of popularity by artificially generating a high volume of retweets for their posts. In our work, we observe a thriving ecosystem of spammers, content advertisers, users paying for content promotion, bots disguised as regular users promoting content and humans retweeting for various incentives. Such content is at best vacuous, but often spammy or malicious and detracts from Twitter content's credibility and honest users' experiences.

Despite previous efforts on Twitter fraudsters' activity [8, 19, 18], the different manifestations of fake retweets have not been adequately studied. Previous approaches focus mainly on specific URL broadcasting, instead of retweet threads, and rely on temporal and textual features to identify bots [5, 12]. Fraudsters on Twitter, though, constantly evolve and adopt advanced techniques to obscure their activities. The identification of patterns associated with "fake" retweet activity is, thus, crucial for spotting retweet threads and their authors as fraudulent. This work's primary goal is to distinguish organic from fake retweet activity and the informal problem definition we address is

Informal Problem 1 (RETWEET-THREAD LEVEL)**.**
Given*: the connectivity network (who-follows-whom); the i-th tweet of user; and the retweet activity (IDs and timestamps of the users that retweeted it)*
Find*: features of the retweet activity*
To determine *whether the activity is organic or not.*

Here, we focus on identifying features and patterns in relation to the connectivity and temporal behavior of retweeters that will allow the classification of the motive behind retweet threads as driven by users' genuine reactions to tweeted content, or resulting from a paid contract. We also aim at spotting users who are suspicious of long-term spam activity, but manage to evade suspension from Twitter by using *camouflage*.



(a) honest user MP 1 (b) honest user HP 1 (c) fraudulent user FD 1

Fig. 1: CONNECTIVITY: Retweeter networks for retweet threads of size (a) 117, (b) 1132, (c) 336. Dense connections in (c) indicate the TRIANGLES pattern. Retweeter networks of honest and fake activities can be distinguished by several other patterns (e.g. DEGREES, HOMOGENEITY). In the depicted networks, a double edge indicates a reciprocal relation, whereas a node's size is relative to its degree.

The contributions of this work are the following:

- **Patterns:** We discover multiple patterns indicative of fraudulent behavior by analyzing the retweeter networks of Twitter accounts. For example, in one class of fraudulent accounts, all accounts follow each other and thus have an excessive number of triangles ("TRIANGLES" pattern) — see Figure 1. It is important that these patterns can be detected based on partial snapshots of the fraudsters' relationship network. Moreover, other fraudsters retweet concurrently within a fixed time from each-other in lockstep fashion, with little variation ("HOMOGENEITY" pattern).
- **Generator:** Based on our analysis, we provide RTGEN, a data generator which produces (ID, timestamp) pairs mimicking traces of fraudulent as well as organic retweet activity. The significance of RTGEN is highlighted by the difficulty of obtaining real world organic and fraudulent retweeting data for experimentation, due to the lack of a standard dataset and the strict policies of social network APIs.
- **Reproducibility:**We share an (anonymized) version of our dataset and RTGEN's code at: `https://app.box.com/s/32pq2bsbtz7nn832cmjm`.

The paper's remainder is organized in a typical fashion — next sections discuss related work, background, the dataset, our proposed method, discoveries and conclusions.

2 Related Work

Related work mainly spans: *anomaly detection* in social networks and *fraud* on Twitter.

Anomaly detection and fraud detection in social networks has led to several methods: NetProbe [14] identifies fraud on eBay using belief propagation. MalSpot [13] uses tensor decomposition for computer network intrusion detection. CopyCatch [1] spots lockstep behavior in Facebook Page Like patterns. [6] leverages spectral analysis to reveal various types of lockstep behavior in social networks.

Fraud on Twitter: [19] analyzes the relationships of criminal accounts inside and outside of the criminal account community to infer types of accounts which serve as criminal supporters. [2] proposes a classification method relying on tweeting behavior, tweet content and account properties for computing the likelihood of an unknown user being a human, bot or cyborg. [17] shows the strong classification and prediction performance of temporal features for distinguishing between account types. All these works, though, address the detection of spammers based on their tweeting and/or networking activity, instead of the fake retweeting problem. In addition, most existing methods (e.g. [18]) consider the typical and out-dated model of a fraudster who has uniform posting frequency and a followers-to-followees ratio close to 1 — nowadays, many fraudsters are more sophisticated. [5] addresses a problem similar to ours, but uses the URLs found in tweets instead of retweet threads in conjunction with a time and user-based entropy to classify posting activity and content. [9] applies *disparity*, also known as *inverse participation ratio* [3], on Twitter data to reveal favoritism in retweets. Table 1 outlines the characteristics of existing methods compared to RTScope, our proposed approach for retweet fraud detection.

Table 1: RTScope comparison against alternatives

	[5]	[19]	[2]	[17]	RTScope
Can be applied for individual retweet chains	✓				✓
Can operate without timestamps		✓			✓
Independent of tweet content	✓			✓	✓
Exploits network topology		✓			✓
Detects bot activity	✓		✓	✓	✓

3 Background on Fake Retweet Thread Detection

Our intitial intuition is that a large proportion of “fake” retweets originate from bot accounts or human accounts which employ the use of automated software. This implies the existence of similarity in the temporal behavior of the individual retweeters, due to the posting (and retweeting) scheduling capabilities of automation tools. We also expect that it is highly probable that fraudulent retweeters of a given user will operate concurrently in lockstep fashion. This is indicative of collaboration between spammers or a contract between the author and a third party for a purchase of retweets. To study the retweeting activity in terms of time and retweeting users, given a user u_m (*author*) we represent the i^{th} tweet posted by u_m with $tw_{m,i}$ as a tuple $(u_m, t_{m,i})$, where $t_{m,i}$ is the tweet’s creation time. Then, a retweet thread is defined as follows:

Definition 1 (Retweet thread). *Given an author u_m and a tweet $tw_{m,i}$, a retweet thread $R_{m,i}$ is defined as the set of all tweets that retweeted $tw_{m,i}$.*

We hypothesize that certain types of fraudulent retweet threads are generated by users with abnormal connectivity in terms of their *follow* relationships in Twitter. An example of such *abnormal connectivity* would be a much denser network of fraudulent (compared to honest) retweeters, corresponding to a group of fraudsters following each other in an attempt to maintain reputability. To validate our hypothesis on the importance of connectivity as a feature, we consider the following two types of relationship networks:

Definition 2 (Relationship networks). *Given a retweet thread $R_{m,i}$ we define the “R-A” and “R” networks as the induced networks of:*
“R-A” network *author u_m and all retweeters of $tw_{m,i}$;*
“R” network *all retweeters of $tw_{m,i}$ minus zero-degree nodes, i.e. retweeters who are disconnected from the rest.*

We highlight the fact that the considered network types are partial snapshots of the complete Twitter followers network, since we operate under the constraint of limited visibility. Constraining the followers network to specific subgraphs is important given that the massive size of the Twitter network poses computational burdens to the application of graph algorithms for pattern detection.

We then formulate two versions of the fake retweet detection problem.

*Problem 1 (*RETWEET-THREAD LEVEL*).*

Given: a tweet $tw_{m,i}$ and a retweet thread $R_{m,i}$,
Identify: whether $R_{m,i}$ is organic.

*Problem 2 (*USER LEVEL*).*

Given: a user u_m, a set of tweets $tw_{m,i}$ and their induced retweet threads,
Identify: whether u_m is a spammer.

The RETWEET-THREAD LEVEL problem addresses the detection of single instances of fraud, thus is suitable for "occasional" fraudsters (who occasionally purchase retweets or are paid to participate in promotions, but otherwise exhibit normal activity) and *promiscuous* professional spammers (their fake retweet threads can be spotted without additional data on their past activities). The USER LEVEL problem addresses also the detection of more *cautious* spammers, whose retweet threads are not suspicious on their own, but thet reveal suspicious recurring patterns when they are jointly analyzed.

4 Dataset and Preliminary Observations

We examine our hypotheses on dataset comprising several retweet threads of honest and fraudulent Twitter users. *method* requires that *complete* retweet threads, i.e. with no gaps in the tuples representing a given tweet's retweets. Due to Twitter Streaming API's constraint of allowing access to only a sample of the published posts, our requirement for *complete* retweet threads, and the lack of a relevant (labeled) dataset, we manually selected a set of *target* users, so that we could track all their posts and retweets for a given time period.

Table 2: Activity statistics per user class

Type	# Tweets	# Original tweets	# Retweeted tweets	# Retweets
honest	35,179	18,706	13,261	708,814
fraudulent	92,520	50,536	27,809	5,330,407
BOTH	127,699	69,242	41,070	6,039,221

We selected target user accounts based on two approaches. The first involved the examination of a 2-day sample of the Twitter timeline, followed by the identification of the users who had posted the most retweeted tweets, and those who posted tweets containing keywords heavily used in spam campaigns (e.g. casino, followback). The second approach was based on "Twitter Counter"[1], a web application publishing lists that rank Twitter users based on criteria such as their number of followers and tweets, and involved the selection of users based on their posting frequency and influence (i.e. we kept only users who posted several posts per week and had received more than 100 retweets on some of their recent posts). We manually labeled target users as "fraudulent" (FD) if (a) inspection of their tweets' content led to the discovery of spammy

[1] http://twittercounter.com/

links to external web pages, spam-related terms, and repetitive posts with the same promotions, or (b) their profile information was clearly fabricated. We labelled the rest of target users (of different popularity scales for the sake of diversity) as "honest" and further divided them into high-, medium- and low-popularity (HP, MP, LP, respectively), using the cut-offs of >100K followers for HP and $< 10K$ followers for LP. We monitored the initial set of target users for 30 days and eliminated those who had all their posts retweeted less than 50 times. Then, we reinforced the remaining dataset with an extra number of similarly selected users, and collected data for an additional 60-days period. At the end of this period, we again pruned users using the same filtering criterion. Overall, this process left a total number of 24 users in the dataset, of which 11 honest (5 HP, 4 MP, and 2 LP) and 13 fraudulent, while after the end of the monitoring period we identified that 4 of our fraudulent users had been suspended by Twitter. Table 2 shows the activity characteristics for the dataset's honest and fraudulent users. For the reproducibility of our results, we make available an anonymized version of our dataset at `https://app.box.com/s/32pq2bsbtz7nn832cmjm`, where readers can also find more detailed information on the activity characteristics for each class of users.

From our data collection and preliminary analysis, we make two main observations:

Observation 1 (Variety)**.** *Fraudsters have various behaviors in terms of their posting frequency and timing.*

Specifically, some fraudsters are *hyperactive*, posting many tweets (> 100 per day); others are more *subtle*, posting few tweets per day, while sometimes mixing original posts with retweets to other users' posts, implying some type of cooperation (half of our dataset's FD users are *hyperactive*). We also noticed that some FD users often produced (resembling) honest posts along with fraudulent ones. This may indicate the existence of "occasional" fraudsters, or intended *camouflage* practiced by "professional" fraudsters.

Observation 2 (FF imbalance)**.** *Despite earlier reports of success, the followers-to-followees ratio (FF) is uninformative for several fraudsters.*

The reasoning behind this observation is that although previous works considered fraudsters with a similar number of followers and followees, we found that some fraudsters maintain a high FF ratio (in our dataset, only two FD users have a ratio close to 1, while for the rest it ranges in 1.3 - 2061). Further complicating the problem, hijacked accounts have honest followers and followees with normal FF ratio.

Given the various types of fraudulent behavior types and inefficacy of the commonly used FF ratio, what additional features can we use to spot fake retweets? This is exactly the focus of RTSCOPE, which is described next.

5 RTScope: Discovery of Retweeting Activity Patterns

In this section we propose RTSCOPE, an approach for detecting fake retweet threads in Twitter and present the results of its application on our dataset. RTSCOPE includes a series of tests that address:

- the RETWEET-THREAD LEVEL problem (1), namely: *ConR*, connectivity analysis of "R" and "R-A" relationship networks (Sect. 5.1);
- the USER LEVEL problem (2), namely: *RAct*, detection of retweeters' activation patterns across a given user's posts (Sect. 5.2), and *ASum*, inspection of the activity summarization features per retweet thread (Sect. 5.3).

The most significant features involved in each test are summarized in Table 3. We note here that in this approach only the *ASum* features require the retweets' timestamps, which, in some cases, may be hard to obtain, or easy for the fraudsters to manipulate.

Table 3: Signs and explanations of suspicious retweeting activity

Feature Category	**Alias**	**Description**	**Fraud Sign**
RETWEET-THREAD LEVEL			
Retweeters' connectivity	*ConR1*	Number of triangles (TRIANGLES)	Excessive
	ConR2	Distribution of degrees (DEGREES)	Non power-law
Activity summarization features	*ASum1*	Activated followers ratio (ENTHUSIASM)	High
	ASum2	IQR (=spread) of interarrival times (MACHINE-GUN)	Low
USER LEVEL			
Retweeters' activation pattern	*RAct*	Distr. of # retweets (HOMOGENEITY)	Homogeneous
Activity summarization features	*ASum3*	Formation of microclusters (REPETITION)	Yes

5.1 Retweeter Networks Connectivity: TRIANGLES & DEGREES Patterns

To study the connectivity between the retweeters of a given tweet, we selected a sample of the largest retweet threads for each user in the dataset, identified their follower relations via the Twitter API and generated the "R" and "R-A" graphs[2]. Interestingly, we observed that for some retweet threads of fraudulent users there were no connections between the retweeters, whereas for others, none of the retweeters was connected to the author. These phenomena were mostly observed in the context of *occasional fraudsters*. However, we noticed that in these cases, a significant (more than 20%) percentage of the original retweeters were suspended some time afterwards, thus affecting the remaining users' connectivity. For the rest of the retweet threads (of fraudulent and honest users) the percentage of suspended retweeters was less than 10%.

The connectivity analysis of the "R" and "R-A" networks led to Observation 3. Next, we discuss the details of our analysis approach and findings.

Observation 3 (CONNECTIVITY)**.** *"R" and "R-A" networks of honest and fraudulent users differ substantially and exhibit the* TRIANGLES*,* DEGREES *and* SATELLITE *patterns, on which we elaborate below:*

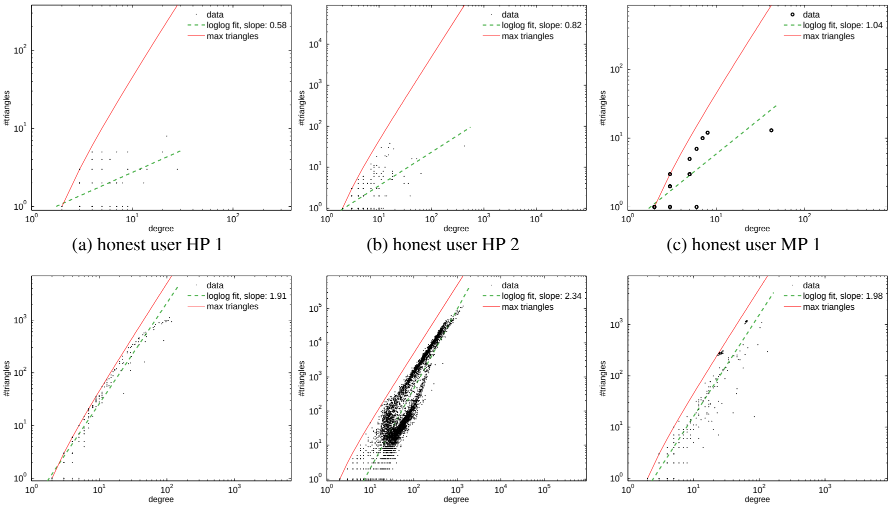
TRIANGLES: Some fraudulent users have a very well connected network of retweeters, resulting in many triangles in their "R" network. The triangles vs. degree plots of fraudsters often exhibit power-law behavior with high (1.1-2.5) slope. Figure 2 shows that honest users (top row, (a)-(c)) have "R" networks with <100 and often 0 triangles. Conversely, the "R" networks of fraudulent users (bottom row, (d)-(f)) are near-cliques with almost the maximum count of triangles for each node ($(d-1)(d-2)/2$ for a node of degree d).

Such networks are probably due to several bot accounts created by a script and made to follow each other in botnet fashion.

DEGREES: Honest users have "R-A" and "R" networks with power-law degree distribution (Figure 3(a)) while fraudulent ones deviate (Figure 3(b)). The spike at degree ≈ 30 for the latter, agrees with the botnet hypothesis.

SATELLITE: In honest "R-A" networks, the author has many "satellites," or retweeters that follow him, and no other retweeters. The fraction s of such satellite nodes is $0.1 < s < 0.9$ for honest users, but $s < 0.001$ for many fraudulent users.

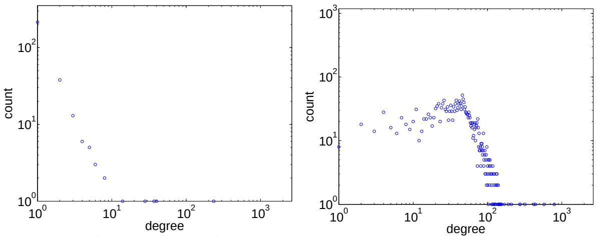
[2] Due to the hard limits of Twitter API in terms of requesting information on users' relations, it was impossible to generate the "R" networks for all retweet threads of the dataset.



(a) honest user HP 1 (b) honest user HP 2 (c) honest user MP 1

(d) fraudulent user FD 1 (e) fraudulent user FD 2 (f) fraudulent user FD 3

Fig. 2: **Dense "R" networks for fraudsters** (TRIANGLES pattern): log-log scatter plots of the number of triangles vs. degree, for each node of selected users' "R" networks. Red line indicates maximum number of triangles ($\approx$ degree^2 for a clique). Dashed green line denotes the least squares fit. Honest users (top) have fewer triangles and smaller slope than fraudsters (bottom).



(a) honest user HP 3 (b) fraudulent user FD 4

Fig. 3: **Fraudsters disobey the degree power-law** (DEGREES pattern): log-log scatter plots of count of nodes with degree deg_i vs. degree deg_i for "R" networks of selected users. Honest users, as depicted in (a), tend to follow power-law behavior whereas fraudsters, as depicted in (b), do not.

5.2 Retweet Activity Frequency: FAVORITISM & HOMOGENEITY Patterns

Given a target users' posts, what is the distribution of retweets across the retweeters? Do most retweets originate from a specific set of *dedicated* users, or are they distributed uniformely across all the user's connections?

To investigate this distibution, we use the *disparity* measure which quantifies, given a finite number of instances (in our case, retweets), the number of different states or subsets these instances can be distributed into. With respect to a given target user, the number of instances corresponds to the total number of retweets, while a given state is the number of retweets made by a single user. Disparity reveals whether the retweeting activity spreads *homogeneously* over a set of users, or if it is strongly *heterogeneous*, in the sense that is skewed towards a small set of very active *dedicated* retweeters.

Given target user u_i and a retweet thread size of k which has been generated by u_j for $j = 1 \ldots k$ retweeters, we examine disparity with respect to the total retweeting activity of these k users. We define the number of retweets made from user j to user i as r_{ij}, and the total number of retweets made from u_j users as $SR = \sum_{j=1}^{k} r_{ij}$. Then, we can consider that the number of retweets r_{ij} defines the *state* of user u_j, which ranges from $r_{ij} = 1$ to $r_{ij} = SR$.

Definition 3 (Disparity). *The disparity of retweeting activity with respect to author u_i and a retweet thread size k is defined as:*

$$Y(k,i) = \sum_{j=1}^{k} (\frac{r_{ij}}{SR})^2 \tag{1}$$

In the case that there exists more than one retweet thread of size k, we simply take the average of the $Y(k,i)$ values over retweet threads.

To give an intuition of diparity, we provide two extreme examples of activity distribution: (a) the *homogeneous*, where all users are in the same state (i.e. they have the same r_{ij} value), and (b) the *super-skewed*, where there exists some user u_l who is at a state of much larger value compared to the rest of the users — that is, $r_{il} \simeq SR$, whereas for $j \neq l$, $r_{ij} = q << SR$. The disparities for these situations are derived as follows:

Lemma 1. *The disparity $Y_{homogeneous}(k,i)$ for the homogeneous activity distribution obeys*

$$Y_{homogeneous}(k,i) = \sum_{j=1}^{k} (\frac{r_{ij}}{SR})^2 = \sum_{j=1}^{k} (\frac{1}{k})^2 = \frac{1}{k} \tag{2}$$

Lemma 2. *The disparity for the super-skewed activity distribution is given by:*

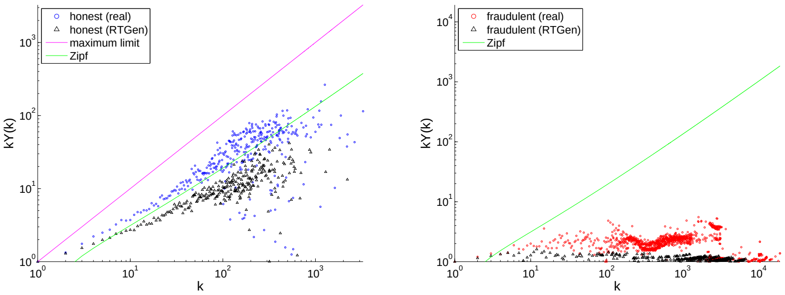
$$Y_{super-skewed}(k,i) = \sum_{j=1}^{k} (\frac{r_{ij}}{SR})^2 = (\frac{r_{il}}{SR})^2 + \sum_{j,j\neq l} (\frac{b}{SR})^2 \simeq 1 \tag{3}$$

, thus it is independent of the retweet thread's size k.

Figure 4 exhibits the relation between $Y(k,i)$ and k averaged over all honest (Figure 4a) and fraudulent users (Figure 4b). We observe that $kY(k,i)$ for honest users appears to have an exponential relationship with respect to k, with an exponent of less than 1 (from equation 3). Fraudulent users' activity is fundamentally different and is close to the homogeneous case, where $kY(k,i) = 1$. The most homogeneous behavior is encountered at large values of k which correspond to heavily promoted tweets, whereas less homogeneity is encountered for small retweet threads, likely for camouflage-related reasons.

We try to approximate the relationship between disparity and k under the hypothesis that the different states r_{ij} of users u_j for $j = 1 \ldots k$ follow a Zipf distribution. If we sort the different r_{ij} states by decreasing order of magnitude, we can express the j^{th} frequency $p_j = \frac{r_{ij}}{SR}$ as $p_j = \frac{1}{j \times \ln{(1.78 * k)}}$[16]. Then, we derive the following lemma:

Lemma 3. *The disparity of a Zipf distribution is given by: $Y_{Zipf}(k,i) \simeq \frac{k-1}{k \times \ln^2{(1.78 * k)}}$*



(a) honest users

(b) fraudulent users

Fig. 4: **Fraudsters exhibit uniform retweet disparity.** (FAVORITISM and HOMOGENEITY patterns): log-log scatter plots of $kY(k,i)$ vs. k for real and simulated retweets of (a) honest users and (b) fraudulent users. Magenta (green) line corresponds to the *super-skewed* case of eq. 3 (the realistic Zipf distribution of Lemma 3). Black triangles correspond to RTGEN retweet threads for: honest-like, in (a) and fraudulent-like, in (b).

Proof. As per equation 1, the disparity of the Zipf distribution can be approximated by:

$$\begin{aligned} Y_{Zipf}(k,i) &\simeq \int_{j=1}^{k} (\frac{1}{j \times \ln{(1.78 * k)}})^2 \\ = \frac{1}{\ln^2{(1.78 * k)}} \int_{j=1}^{k} \frac{1}{j^2} &= \frac{k-1}{k \times \ln^2{(1.78 * k)}} \square \end{aligned} \tag{4}$$

Figure 4a depicts the k-$kY_{Zipf}(k,i)$ relation with a green line, which is a good fit for honest users' behavior (FAVORITISM pattern). Conversely, fraudulent users' disparity is characteristic of a zero slope (HOMOGENEITY pattern), as indicated by Figure 4b.

Observation 4 (FAVORITISM)**.** *The disparity of retweeting activity to honest users' posts can be modeled under the hypothesis that the participation of users to retweets follows a Zipf law.*

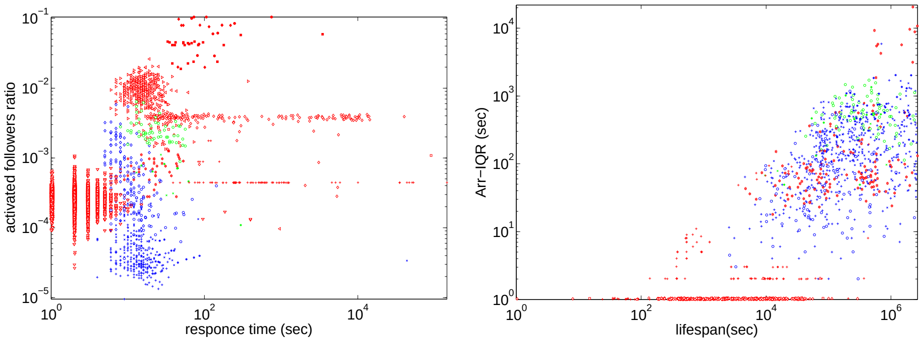
Observation 5 (HOMOGENEITY)**.** *The disparity of retweeting activity to fraudulent users' posts can be modeled under the hypothesis that the participation of users to retweets is homogeneous.*

5.3 Activity Summarization Features: MACHINE-GUN, ENTHUSIASM & REPETITION Patterns

We further extracted the following temporal and popularity features with respect to the retweet threads included in the datasets:

- *ratio of activated followers*, i.e. author's followers who retweeted;
- *response time*, i.e. time elapsed between the tweet's posting and its first retweet;
- *lifespan*, i.e. time elapsed between the first and the last (observed) retweet, constrained to 1 month to remove bias with respect to later tweets;
- *Arr-IQR*, i.e inter-quartile range of interarrival times for retweets.

Figure 5a depicts the scatterplot of activated followers ratio vs. response time for retweet threads of all target users. Interestingly, several red points of users suspected of fraud are clearly separated from honest users' retweet threads due to their high or



(a) Activated followers ratio vs. Response time

(b) Arr-QR vs. Lifespan

Fig. 5: **Dense microclusters formed by fraudsters.** (ENTHUSIASM, MACHINE-GUN and REPETITION patterns): log-log scatter plots of ActSum features for all target users - each point is a retweet thread, each author has a different glyph, HP, MP, LP users are in blue, green, cyan, and fraudsters are in red.

low response time and high *activated followers ratio*. In addition, the consideration of various feature combinations can be useful for identifying fake retweet threads. Figure 5b, which depicts the scatter plot of the Arr-IQR vs. lifespan for retweets of all target users' retweet threads, indicates that several retweet threads of the same fraudulent users tend to exhibit similar values for these features, resulting in the formation of dense microclusters of points. For example, the cluster appearing at the figure's bottom-left side is created from retweet threads whose author is fraudulent user FD 5.

From this analysis, we draw several additional observations.

Observation 6 (ENTHUSIASM)**.** *Followers of fraudulent retweeters have a high infection probability.*

Observation 7 (MACHINE-GUN)**.** *Fraudsters retweet all at once, or with a similar time-delay.*

Observation 8 (REPETITION)**.** *Groups of fake retweet threads exhibit the same values in terms of response time, Arr-IQR and activated followers ratio, forming microclusters.*

6 RTGen Generator

We propose RTGEN, a generator that simulates the retweeting activity of honest and fraudulent users, highlight its properties, and present its results with respect to disparity.

Algorithm 1 outlines the process for the simulation of the retweeting behavior over a network $G(V, E)$, where V_i is the set of users and $E_{i,j}$ is the set of directed *who-follows-whom* relationships between them. In our model, a given user u_i from the set V_i is considered a *candidate* for retweeting if u_i follows either the *author* or another user who has already retweeted (an activated user). Each run of the generator involves the selection of a random user and the simulation of the tweet forwarding process for N tweet events. More specifically, in the first simulation, the *author* of a tweet is randomly selected, and the author's followers become candidate retweeters. Each candidate is then added to a list of activated users with a given retweeting probability. This process is executed recursively until all activated users' followers have been examined and there are no more candidate users. Then, RTGEN continues with the next simulation. Each simulation (tweet) is characterized by a varying *interestingness* value representing the infection probability given the significance of the tweet's content.

```
**Data**: $G(V, E)$ = Examined network, $N$ = number of simulations, $b$ = interestingness in
    $[B_1,..., B_n]$
**Result**: $activatedUsers$ : activated nodes $\in V$ per simulation
$author \leftarrow$ user randomly selected from $V$;
$sim \leftarrow 1$ ;
while $sim \leq N$ do
    $initialInterestingness \leftarrow$ pick an interestingness $b$ from $B_i$ ;
    $candidateUsers \leftarrow$ authors' followers ;
    for each user in candidateUsers do
        $followers \leftarrow$ take followers of candidateUsers ;
        for each follower $f$ in $followers$ do
            if $f$ not in activated users then
                calculate retweet probability $bUser_f$ ;
                add $f$ to $activatedUsers$ with probability $bUser_f$;
    $sim \leftarrow sim + 1$ ;
```

Algorithm 1: Pseudocode for RTGEN

RTGEN simulates the scenarios of honest and fraudulent retweeting behavior by forming hypotheses on the underlying graph and the users' inclination to retweet. In specific, based on the discovered TRIANGLES and DEGREES patterns, RTGEN uses a Kronecker graph [11] to simulate honest users' networks and a dense Erdös-Rényi graph [4] for fraudsters' networks. Moreover, RTGEN assumes the same infection probability for all fraudulent users, based on the ENTHUSIASM and REPETITION patterns. Conversely, honest users have different activation rates depending on the tweet's interestingness, topics of interest and limited attention. For generality, we follow the *weighted cascade model* [7] and assume that user u_i's infection probability is inversely proportional to the number of followers. This lowers the retweeting probability for users with a large number of followers, simulating limited attention and content competition. For organic retweet thread simulation, the probability $bUser_v$ of user v is thus taken as:

$$P_{honest}(v, i) = b_i * (1/|f_v|) \tag{5}$$

where $b_i \in [B_1, ..., B_n]$ is the tweet's interestigness in the i_{th} simulation sim_i and $|f_v|$ is the number of followers for user v. Respectively, for the fake retweet thread case:

$$P_{fraudulent}(v, i) = b_i \tag{6}$$

where, here, b_i is randomly selected between two probability values [B_1, B_2]. B_1 represents *camouflage* retweeting activity, and B_2 represents *fake* retweeting activity, with B_2 being much higher than B_2 (in our experiments by an order of magnitude).

RTGEN was applied on: (a) a Kronecker graph of 500k nodes, 14M edges (generated with a parameter matrix $\begin{pmatrix} 0.9999 & 0.5542 \\ 0.5785 & 0.2534 \end{pmatrix}$ [15]), and (b) an Erdös-Rényi graph of 10k nodes, 1M edges, for 10 users and 100 simulations. Based on the simulation results, we calculated the disparity for each author and k-sized retweet thread and averaged the disparity values separately for honest and fraudulent authors. Figure 4 depicts the relation between disparity and k for each class of users, which emulate those derived from real Twitter data.

7 Conclusions

Fake retweet behavior incentivized by monetary and social benefits negatively impacts the credibility of content and the perception of honest users on Twitter. In this work, we focus on spotting fake from organic retweet behavior, as well as identifying the fraudsters to blame by carefully extracting features from the activity of their retweeters. Specifically, our main contributions are:

- **Patterns:** We discovered several patterns for characterizing various types of fraudulent users: e.g. the "TRIANGLES" pattern revealing strong connectivity in retweeter networks, the "HOMOGENEITY" pattern indicating uniform retweet disparity.
- **Generator:** We propose RTGEN, a scalable, realistic generator which produces both organic and fraudulent retweet activity using the weighted cascade model. RTGEN can be useful for experimentation and evaluation scenarios where actual, labeled retweet data are missing.

Reproducibility: We share both our retweet thread dataset and RTGEN's code.

References

1. A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos. Copycatch: stopping group attacks by spotting lockstep behavior in social networks. In *WWW*, 119–130. ACM, 2013.
2. Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? *ACSAC*, 21–30, 2010.
3. B. Derrida and H. Flyvbjerg. Statistical properties of randomly broken objects and of multivalley structures in disordered systems. *Journal of Physics A: Mathematical and General*, 20(15):5273–5288, 1987.
4. P. Erdos and A. Renyi. On the evolution of random graphs. *Publ. Math. Inst. Hungary. Acad. Sci.*, 5:17–61, 1960.
5. R. Ghosh, T. Surachawala, and K. Lerman. Entropy-based classification of retweeting activity on twitter. In *KDD workshop on Social Network Analysis (SNA-KDD)*, Aug. 2011.
6. M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Inferring strange behavior from connectivity pattern in social networks. *PAKDD*, May 13 - May 16 2014.
7. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, 137–146, New York, NY, USA, 2003. ACM.
8. T. Kurt, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of twitter spam. *IMC*, 243–258, 2011.
9. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? *WWW*, 591–60, 2010.
10. K. Lee, B. Eoff, and J. Caverlee. Seven months with the devils: A long-term study of content polluters on twitter. *ICWSM*, 2011.
11. J. Leskovec, D. Chakrabarti, J. M. Kleinberg, C. Faloutsos, and Z. Ghahramani. Kronecker graphs: An approach to modeling networks. *Journal of Machine Learning Research*, 11:985–1042, 2010.
12. P.-C. Lin and P.-M. Huang. A study of effective features for detecting long-surviving twitter spam accounts. *ICACT*, 841, 27-30 Jan. 2013.
13. C.-H. Mao, C.-J. Wu, E. E. Papalexakis, C. Faloutsos, and T.-C. Kao. Malspot: Multi^2 malicious network behavior patterns analysis. In *PAKDD*, 2014.
14. S. Pandit, D. H. Chau, S. Wang, and C. Faloutsos. Netprobe: a fast and scalable system for fraud detection in online auction networks. In *WWW*, 201–210. ACM, 2007.
15. A. Rao, S. Sripada, and G. K. Parai. Modeling and analysis of real world networks using kronecker graphs. Project report, 2010.
16. M. Schroeder. *Fractals, Chaos, Power Laws*. W. H. Freeman, New York, 6 edition, 1991.
17. G. Tavares and F. A. Scaling-laws of human broadcast communication enable distinction between human, corporate and robot twitter users. *PLoS ONE*, 8(7):e65774, 2013.
18. X. Wu, Z. Feng, W. Fan, J. Gao, and Y. Yu. Detecting marionette microblog users for improved information credibility. *Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science*, 8190:483–498, 2013.
19. C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. *WWW*, 71–80, 2012.