Retweet Us, We Will Retweet You: Spotting Collusive Retweeters Involved in Blackmarket Services

Hridoy Sankar Dutta
IIIT Delhi, India
hridoyd@iiitd.ac.in

Aditya Chetan*
IIIT Delhi, India
aditya16217@iiitd.ac.in

Brihi Joshi*
IIIT Delhi, India
brihi16142@iiitd.ac.in

Tanmoy Chakraborty
IIIT Delhi, India
tanmoy@iiitd.ac.in

***Abstract*—Twitter has increasingly become a popular platform to share news and user opinion. A tweet is considered to be important if it receives high number of affirmative reactions from other Twitter users via `Retweets`. *Retweet count* is thus considered as a surrogate measure for positive crowd-sourced reactions – high number of retweets of a tweet not only help the tweet being broadcasted, but also aid in making its topic trending. This in turn bolsters the social reputation of the author of the tweet. Since social reputation/impact of users/tweets influences many decisions (such as promoting brands, advertisement etc.), several blackmarket syndicates have actively been engaged in producing fake retweets in a collusive manner. Users who want to boost the impact of their tweets approach the blackmarket services, and gain retweets for their own tweets by retweeting other customers' tweets. Thus they become customers of blackmarket syndicates and engage in fake activities. Interestingly, these customers are neither bots, nor even fake users – they are usually normal human beings; they *express a mix of organic and inorganic retweeting activities*, and *there is no synchronicity across their behaviors*.**

In this paper, we make a first attempt to investigate such blackmarket customers engaged in producing fake retweets. We collected and annotated a novel dataset comprising of customers of many blackmarket services and characterize them using a set of 64 novel features. We show how their social behavior differs from genuine users. We then use state-of-the-art supervised models to detect three types of customers (bots, promotional, normal) and genuine users. We achieve a Macro F1-score of 0.87 with SVM, outperforming four other baselines significantly. We further design a browser extension, SCoRe which, given the link of a tweet, spots its fake retweeters in real-time. We also collected users' feedback on the performance of SCoRe and obtained 85% accuracy.

***Index Terms*—Retweeters, collusion, blackmarket, Twitter, Online Social Networks.**

I. Introduction

Twitter, arguably the most popular micro-blogging site, provides its users two major ways to place their affirmative reactions towards different entities: (i) user-level affirmation (such as `follow`), and (ii) content-level affirmation (such as `retweet`, `like`). Here, we particularly focus on 'retweeting activity', a major content-level affirmative action, which provides a way of re-broadcasting messages and confirms retweeter's agreement of the message being important to others. Retweet count of a tweet gives the OSN users a sense of crowdsourced agreement on the tweet, and thus determines the influence of the tweet as well as the author of the tweet. It also helps in making a certain topic trending on Twitter. This further brings an opportunity of 'intentional manipulation' to the social adversaries who falsely create the impression of popularity of tweets by generating huge volume of retweets.

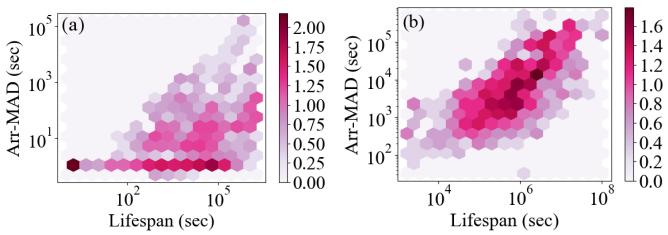

Fig. 1: **(b) Asynchronous behavior of collusive retweeters as opposed to the (a) synchronicity of normal retweet fraudsters mentioned in [10]**. The figure shows two binned 2-D heatmaps (in logarithmic scale) of *Lifespan* (time elapsed between the first and last retweets) vs. *Arr-MAD* (mean absolute deviation of retweets inter-arrival times) of retweet threads for (a) normal retweet fraudsters (data taken from [10]) and (b) collusive retweeters. Unlike (a) where *Arr-MAD* is almost invariant with *Lifespan*, (b) shows an increasing trend.

TABLE I: **Macro F1-score of the competing methods.** None of the existing methods can detect collusive retweeters accurately.

Bot Detection [5]	Fake Account Detection [8]	Sync. Fake Retweeter Detection [10]	Our Method
0.760	0.693	0.750	0.873

Challenges in detecting collusive retweeters: There exist several blackmarket agencies which have created thriving and intelligent ecosystems of producing spam retweets. Users can gain retweet count of their own tweets for free by retweeting tweets of other customers of those services. Thus users unwillingly become customers of blackmarket services. Such trend of inflating social reputation has become prevalent across different social media platforms. Detection of such blackmarket retweeters is challenging for many reasons – (i) They are not bots, but human beings; therefore, bot detection algorithms can not flag them (first column of Table I). (ii) Their Twitter accounts are not fake; therefore, fake account detection algorithms can not detect them (second column of Table I). (iii) They express a *mix of organic and inorganic behavior* in their retweeting patterns – they organically reweet

*Equal contribution

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

some genuine tweets; at the same time, they inorganically retweet tweets submitted to blackmarket services. The extent of inorganic activities may differ across retweeters. (iv) They do not show any synchronicity across their retweeting patterns, thus making it difficult for the existing synchronous fake retweeter detection methods [10] to detect them (third column of Table I).

Motivation and state-of-the-art: Despite previous efforts in understanding different fraudulent activities in Twitter such as fake account detection [8], social spam detection [23], content affirmation via retweeting has hardly been studied. A preliminarily attempt was made by [10] to detect spam retweeters with the hypothesis that suspicious activities are highly 'synchronized', i.e., a group of spammers retweet at the same time. However, we clearly observe that **synchronicity does not hold among collusive retweeters** (Figure 1) – this is intuitive since the blackmarket services do not have control on the customers involved in producing fraud retweets. A very recent effort was made by [1] to detect collusive followers in Twitter. They concluded that crowdsourced manipulation is a big threat to the credibility of OSNs. **However, to the best of our knowledge, ours is the first attempt to detect blackmarket-based collusive manipulation of content credibility via fraudulent retweeting activities.** Interestingly, Twitter has not yet been successful in flagging these customers – we observed 31 blackmarket customers who are marked as 'verified users' by Twitter.

Our Contributions: We begin our experiment by surveying different types of blackmarket services (Section III-A) and collecting customers from multiple blackmarket services (Section III-B). We employed human annotators to label each customer into one of the following categories – *bots*, *promotional* and *normal*. We also collected genuine users who are researchers / experts in machine learning, following the method in [18]. The timeline information of these users (customers and genuine users) were further scraped. These two sets of users thus form the retweeter set which we further analyze. We then propose an exhaustive set of 64 novel features to characterize retweeters (Section IV). We run several state-of-the-art supervised models to classify retweeters into four types – bot, promotional customer, normal customer, genuine user. We observe that Logistic Regression achieves the maximum accuracy, outperforming the baseline methods significantly (Section V-C). Our method (with SVM) turns out to be equally successful in categorizing retweeters into binary classes – customers and genuine (Macro F1-socre of 0.87).

We finally build a browser extension, SCoRe, '**S**potting **Co**llusive **Re**tweers', which, given a retweeter set of a tweet, classifies each retweeter as customer or genuine. The extension allows the end users to judge the quality of the classification SCoRe produces and provide correct labeling, in case it produces wrong result. The user-generated feedback is collected in the back-end and the supervised model is retrained in an incremental way. However, special care has been taken to ignore spurious feedbacks (more details in Section VI). We seek feedback of 25 volunteers on the performance of SCoRe and obtain 85% accuracy. We believe that the widespread use of such extension would help OSN administrators block the customers and measure the true credibility of a tweet based on its genuine retweet count.

Reproducibility: All the codes and processed dataset are available at https://tinyurl.com/y7ruh2o8.

II. Related Work

Detection of fraudsters in OSNs: Many studies have been conducted on detecting frauds in Twitter and other OSNs. [2] detected spammers in Twitter using features generated from tweet content and user social behavior. [15] identified strangers on Twitter, who can be potential retweeters to help effectively propagate intended information within a desired time frame. [11] focused on detecting inorganic retweet behavior and proposed 'RTGEN', a realistic generator that imitates the behaviors of both honest and fraudulent users. [4] used machine learning approach to classify Twitter accounts into 'human', 'bot' and 'cyborg'. [9] discussed the rise of Social Bots on Twitter. [13] used network properties to identify spammers on Twitter. [23] proposed a spam detection technique to identify suspicious users on Twitter. [20] studied multiple spamming techniques, including creating fake Twitter accounts, generating spam URLs, and spam distribution. [14] used synchronous and abnormal behaviors of Twitter followers to detect suspicious following behavior. [12] detected spam campaigns on Twitter.

Detection of blackmarket customers in Twitter: Understanding the dynamics of blackmarket services have recently gained substantial attention among researchers. [6] provided a detailed analysis of blackmarket services and their impact on multiple OSNs. Most of the prior works showed how fake followers in social media help in promoting a certain user [19]. [18] studied multiple types of blackmarket agencies by showing their multifaceted behavior. [21] studied various blackmarkets tied to fraudulent Twitter credentials, monitoring pricing, availability, and fraud perpetrated by these services. [16] tackled the problem of voluntary following activity detection to automatically detect malicious accounts that make profit in the follower markets. [17] performed an analysis of six different underground forums and showed how OSN users are engaged in selling goods and services.

Remarks. Our methodology differs from the existing approaches in many ways: (i) [10] identified fake retweeting behavior in Twitter on the basis of synchronicity. This type of behavior can only be observed if the retweeters are controlled by a centralized authority. However, collusive blackmarket customers are not controlled by anyone. They intentionally retweet other customers' tweets. Therefore, they do not exhibit any synchronized behavior. (ii) Existing fraud detection approaches classify a user into 'fraud' or 'genuine' [18], [10]. Our work classifies users into four categories – 'genuine', 'bots', 'promotional customers' and 'normal customers'. (iii) We are also the first to develop a real-time system that can identify customers and genuine retweeters, given the retweeter list of a tweet.

III. BLACKMARKET SERVICES

In this section, we describe our efforts in collecting and annotating a set of Twitter accounts corresponding to the customers of various blackmarket agencies.

A. Types of Blackmarket Retweet Services

While investigating the mode of blackmarket services, we noticed that there are two prevalent models of services [18] – (i) ***Premium Services*** which only provide services upon receiving payment from the customers; (ii) ***Freemium Services*** which, similar to premium services, offer both paid services, as well as unpaid services that require the users to provide their Twitter login details; this in turn may involve the users unconsciously in the blackmarket activities. Here our primary focus is to understand the *activities of freemium services* which are easy to access due to their unpaid service model. This type of services can further be divided into three categories:

- **Social-share services:** These services (e.g., FreeFollowers[1]) ask customers to perform activities on social-media contents of other customers involved in those services. The activities on social-media contents can be 'Facebook Share', 'Facebook Like', 'Twitter Retweets' etc.
- **Credit-based services**: These services (e.g., Like4Like[1], YouLikeHits[1], TraffUp[1], JustRetweet[1]) work on credit-based policies. Each customer has to put a value for the tweet, which gets deducted from his/her credits when the tweet is published. A customer retweets the tweets of other customers to earn credits and the value of each tweet is added up to his/her total credit.
- **Auto-time retweet services**: The customers of these services (e.g., TweetsTool[2]) need to get access token from Twitter and login to the service. They can request for 10-50 retweets for each of their tweets in a 15-minute window. Other than the retweet service, it also provides the customers with the following services – 'Auto Favourite', 'Auto Follower', 'Auto Reply'.

B. Data Collection

Collecting blackmarket customers: In order to collect the information of blackmarket customers, we focused our crawling only on credit-based services because – (i) their service policies are easy to understand, (ii) as opposed to the social-share services, they only utilize a single platform (Twitter) to perform their activities, (iii) most of the freemium services follow credit-based strategy compared to the other strategies which would help us in collecting more data for our analysis. We adopted an 'active probing' strategy to collect the dataset – we created multiple dummy accounts in each of the credit-based freemium services (Like4Like, YouLikeHits, TraffUp, JustRetweet), kept retweeting tweets of other customers whose tweets were posted on the blackmarket services, and collected their IDs. We continued these activities for one month (Feb, 2018). Table II shows the statistics of the dataset for each service. During this process, Twitter suspended some of the accounts we collected, and therefore we ignored those accounts for further analysis. From the collected dataset, we noticed that the average number of tweets posted by a user to YouLikeHits is 1.43 which is relatively high compared to Like4Like (1.01). Interestingly, we found a significant number of customer overlap (76 customers) across YouLikeHits and Like4Like. It suggests that customers utilize several such blackmarket services audaciously without any obfuscation of their identities to promote their tweets. We further scraped the timeline information of these customers using Twitter's REST API.

TABLE II: Statistics of the dataset.

Service	# users	# users suspended / deleted	# users taken for our analysis	# tweets our dummy accounts retweeted
YouLikeHits	638	168	470	914
Like4Like	451	178	273	459
TraffUp	2	0	2	6
JustRetweet	11	3	8	1
Genuine Users	1000	0	1000	-

Collecting genuine users: To compare normal users with blackmarket customers in Twitter, we selected 1000 genuine users (following [18]) – (i) We searched for Twitter lists containing users who are researchers, working in the field of Machine Learning/Data Mining/Information Retrieval. (ii) We also discarded genuine users who have more than 1 million followers since high follower count may resemble a celebrity and we wanted to discard any celebrity-like users from our analysis in order to remove any unnecessary bias. We further scraped the timelines of the remaining users.

C. Human Annotation

Three human annotators[3] were asked to label the blackmarket customers into *Bots*, *Promotional Customers* and *Normal Customers* based on our definition of each customer type and Twitter's terms of service. Annotators were also given freedom to search for any information related to customers and apply their own intuition. Here we describe the information we provided to the annotators about three types of customers:

1) **Bots:** A Twitter Bot is a software which controls a Twitter account using the Twitter API [4]. Bots on Twitter perform both helpful and harmful activities working in a coordinated fashion [5].
2) **Promotional customers:** Twitter has always been a source of advertising content to its target users. Promotional users are involved in promoting one of the three products: *tweets*, *accounts* or *services* [7]. The annotators were asked to search for tweets in the timeline of customers promoting one of the above products. We observed that many customers in our dataset were involved in promoting brands using keywords such as 'win', 'ad', 'Giveaway'.

[1]FreeFollowers: https://www.freefollowers.io/, Like4Like: https://like4like.org/, TraffUp: http://traffup.net/, JustRetweet: http://justretweet.com.

[2]http://www.tweetstool.com/

[3]They were experts in social media, and their age ranged between 25-35.

3) **Normal customers:** These customer do not fall under any of the two categories mentioned above.

Each annotator was given all the customer accounts for annotation. Finally, we considered only 743 customers whose labels were agreed upon by at least two annotators, and found 86 bots, 275 promotional customers and 382 normal customers. The average inter-annotator agreement was 0.79 based on Cohen's κ.

D. Interesting Observations

- Customers of the blackmarket services often remove their tweets after a small duration. We found that around 3% customers have removed their tweets.
- We also found 1% customers who are suspended by Twitter. This clearly indicates that Twitter is still inefficient in identifying customers of the blackmarket services.
- We found 31 customers who are marked as verified genuine accounts by Twitter. This also indicates the inability of current Twitter policy to flag blackmarket customers.
- The Twitter profiles of 4% customers are shown with the warning *"Caution: This account is temporarily restricted"*. Twitter enforces this warning on users in regards to either of the following cases : (i) repeatedly posting duplicate or near-duplicate content, (ii) abusing trending topics or hashtags, (iii) sending automated tweets or replies, and (iv) using bots or applications to post similar messages based on keywords. , (v) posting similar messages over multiple accounts, and (vi) aggressively following and un-following people.

IV. Experimental Setup

In this section, we present novel features to characterize different types of users (both customers and genuine users) and supervised models for classifying users.

A. Feature Selection

We use an extensive set of 64 novel features to detect different types of users. We group these features into five buckets: (i) Profile Features (PF), (ii) Social Network Features (SNF), (iii) User Activity Features (UAF), (iv) Likelihood Features (LF), and (v) Fluctuation Features (FF).

(i) Profile Features (PF): We hypothesize that the older the account of a user, s/he will have more tendency to retweet others' tweets. Therefore, we take *account age* (**PF$_1$**) as one feature in our study. Figure 2 corroborates our hypothesis by showing that both genuine users and customers exhibit an increasing trend of retweet count w.r.t their account age. We further consider four other profile features: (**PF$_2$**) *length of the screen name*, (**PF$_3$**) *whether the profile has a description or not*, (**PF$_4$**) *length of profile description*, (**PF$_5$**) *whether profile has a URL or not*.

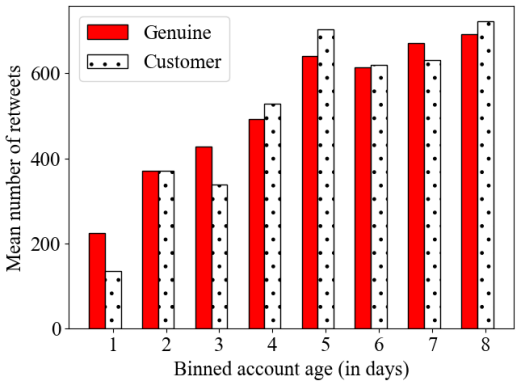

Fig. 2: Relationship between retweet count and account age (in days; further divided into bins: bin 1: 0-499, bin 2: 500-999,$\cdots$, bin 8: 3500-3999).

(ii) Social Network Features (SNF): Social network features such as *number of followees* (**SNF$_1$**) *and followers* (**SNF$_2$**) and *their ratio* (**SNF$_3$**) sometimes provide indicative information about the users. Figure 3(a) shows that customers tend to follow a lot of other users, compared to the genuine users – this may be explained by the typical tendency of customers to follow others with the intention that the followees will further draw attention to their profiles as well as their tweets. Figure 3(a:inset) shows that followee count has a more positive correlation (Spearman's $\rho = 0.58$) with retweet count in case of genuine user than the customers (Spearman's $\rho = 0.42$). Figure 3(b) shows that genuine users tend to have a variety of follower count, whereas all customers have a smaller follower count. Figure 3 (b:inset) shows that genuine users with a large number of followers tend to retweet more compared to those with less followers; however, it may not hold for the customers. The ratio of followees to followers in Figure 3(c) shows that genuine users have a lower followee to follower ratio – this is expected as they tend to have around equal number of followees and followers (recall that we excluded celebrities from our study, whose follower count may be much higher than the followee count). For genuine users, we also observe a declining trend of retweet count with the increase of followee to follower ratio; however customers do not follow any such trend (Figure 3 (c:inset)). We further measure the influence score of users using *Klout score*[4] (**SNF$_4$**) and take it as a feature. Klout score returns a value between 1-100 to rate a user based on his/her online social influence. The Klout score of genuine users and customers based on their influence is shown in Figure 3(d). The median value of Klout score for customers and genuine users is 48.11 and 43.57 respectively – such small difference in Klout score indicates that though customers use blackmarket services, it does not help them making their profile significantly popular. Fig. 3 (d:inset) shows that the retweet count tends to increase with the increase of Klout score for genuine users; however there is not such uniform pattern for customers.

[3]https://maximizesocialbusiness.com/why-your-twitter-account-may-be-restricted-1169/

[4]http://klout.com

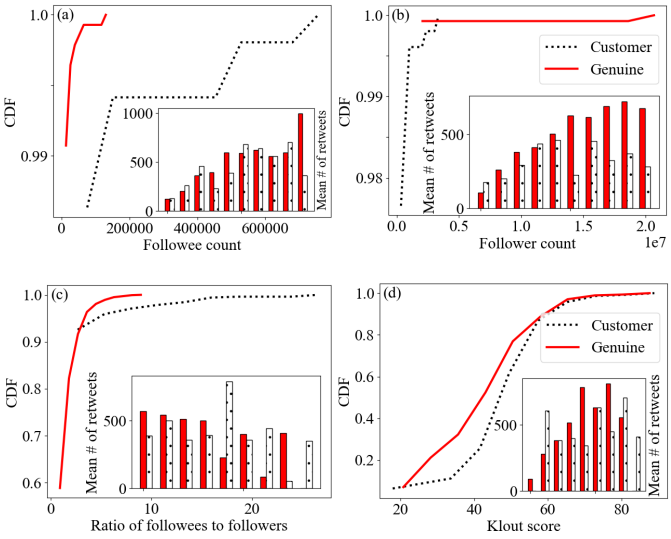

Fig. 3: CDF of social network features – (a) followee count, (b) follower count, (c) followee to follower ratio, (d) Klout score, and their relation with the retweet count (inset). The variable in x-axis is same for the main diagram and its inset.

(iii) User Activity Features (UAF): User activity plays a significant role - higher the activity of a user, higher is the chance of his/her tweet to be retweeted by a stranger. We use the following features to measure the activity of users: (**UAF$_1$**) *total number of tweets*, (**UAF$_2$**) *number of direct mentions per tweet*, (**UAF$_3$**) *number of URLs per tweet*, (**UAF$_4$**) *number of hashtags per tweet*, (**UAF$_5$**) *number of tweets per day*, (**UAF$_6$**) *number of retweets per day*, and (**UAF$_7$**) *number of retweets per tweet*. To compute these features, we crawl all (max 3200) tweets from the timeline of each user. We also use *Bot-score* (**UAF$_8$**) of each user from Botometer service[5] [5] and consider it as a feature.

(iv) Likelihood Features (LF): If Twitter users want to publish tweets in a credit-based blackmarket service, they need to retweet a lot of others' tweets to earn credits for their own tweets to be retweeted by other customers. On the contrary, when they write tweets on Twitter and do not submit these tweets to any blackmarket services, they are not expected to perform such credit-based retweeting activity. We use the following set of features to capture this phenomenon: (**LF$_{1\text{-}7}$**) *tweeting likelihood per day for seven days* (*Monday-Sunday*), (**LF$_{8\text{-}14}$**) *retweeting likelihood per day for seven days*, (**LF$_{15\text{-}21}$**) *regularity of tweeting activity per day for seven days*, (**LF$_{22\text{-}28}$**) *regularity of retweeting activity per day for seven days*, (**LF$_{29}$**) *tweet steadiness*, (**LF$_{30}$**) *retweet steadiness*, (**LF$_{31\text{-}37}$**) *maximum tweet likelihood per day for seven days*, and (**LF$_{38\text{-}44}$**) *maximum retweet likelihood per day for seven days*. **LF$_{1\text{-}7}$** (*resp.* **LF$_{8\text{-}14}$**) is calculated by taking the ratio of the tweets (*resp.* retweets) of a user per day to the total number of tweets (*resp.* retweets) the user posted in a week. Regularity of tweeting activity per day (**LF$_{15\text{-}21}$**) is calculated by $-\sum_{i=1}^{24} p(x_i) \log p(x_i)$, where $p(x_i)$ is the fraction of tweets posted by the user at i^{th} hour of that day. We follow the same method to measure **LF$_{22\text{-}28}$** by replacing tweets with retweets. Tweet (*resp.* retweet) steadiness is calculated by $1/\sigma_t$ (*resp.* $1/\sigma_{rt}$) where σ_t (*resp.* σ_{rt}) is the standard deviation of time difference between consecutive user-generated tweets (*resp.* retweets). **LF$_{31\text{-}37}$** (*resp.* **LF$_{38\text{-}44}$**) is calculated by the ratio of per-day tweet (*resp.* retweet) count of a user to the maximum number of tweets or (*resp.* retweets) the user posted in a day of a week.

(v) Fluctuation Features (FF): Customers of credit-based blackmarket services want to acquire as many credits as possible in order to gain more retweets to their own tweets. We use the following features to validate this phenomenon: (**FF$_1$**) *standard deviation of retweet counts for all user-generated tweets*, (**FF$_2$**) *mean of log-time difference between consecutive retweets*, (**FF$_3$**) *standard deviation of log-time difference between consecutive retweets*.

Note that unlike [3], we did not use any graph-related features for two reasons: (i) crawling the complete neighborhood structure of users requires huge computational resources and sometimes produces incomplete information due to several constraints such as restriction of API and user profiles, (ii) given an unknown user account, collecting its neighborhood structure is time-consuming, which in turn may affect scalability of the real-time system we intended to build (Section VI). However, we do not claim that graph-based features will not enhance the classification performance.

B. Classification Models

We consider six state-of-the-art stand-alone supervised classifiers – Decision Tree (DT), K-Nearest Neighbors (K-NN), Logistic Regression (LR), Naive Bayes (NB), Support Vector Machine (SVM). We also consider three ensemble classifiers: Random Forest (RF), Bagging (BG) and Boosting (BO). We perform hyper-parameter optimization in order to find the parameters that generate the best results. For instance, we use CART with Gini gain criteria for DT; K-NN method with K = 5; multinomial logistic regression and SVM with linear kernel.

V. Experimental Results

In this section, we start by briefly describing four baselines, followed by the detailed experimental results.

A. Baseline Methods

Since there is no prior work on collusive retweeter detection, we choose four state-of-the-art methods as baselines, which are close to the problem we want to solve here.

Baseline I: BotoM**:** We use the social bot detection method proposed by [5] and measure bot score for each retweeter. This score is further used as a feature to run different supervised classifiers mentioned in Section IV-B.

Baseline II: SpamBot**:** We use the method proposed by [22] as our second baseline. It leverages a set of content-based and network-based features to detect spam bots from a collection

[5]https://botometer.iuni.iu.edu

TABLE III: Performance of different competing methods for multi-class classification.

Classifier	Micro				Macro			
	Precision	Recall	F1	ROC-AUC	Precision	Recall	F1	AUC
BotoM (RF)	0.638	0.638	0.638	0.853	0.497	0.596	0.540	0.780
SpamBot	0.565	0.565	0.565	0.658	0.321	0.308	0.287	0.576
FakeAcc (LR)	0.568	0.568	0.568	0.809	0.358	0.383	0.347	0.727
NDSync (LR)	0.573	0.573	0.573	0.764	0.383	0.293	0.269	0.545
Decision Tree	**0.754**	**0.754**	**0.754**	0.706	0.636	0.596	0.587	0.712
K-NN	0.693	0.693	0.693	0.862	0.666	0.519	0.544	0.779
Logistic Regression	**0.754**	**0.754**	**0.754**	**0.924**	**0.731**	**0.641**	**0.671**	**0.909**
Naive Bayes	0.518	0.518	0.518	0.712	0.129	0.250	0.170	0.724
SVM	0.746	0.746	0.746	0.918	0.592	0.558	0.557	0.896
Random Forest	0.668	0.668	0.668	0.863	0.658	0.637	0.639	0.833
Bagging	0.683	0.683	0.683	0.904	0.689	0.707	0.691	0.884
Boosting	0.693	0.693	0.693	0.892	0.666	0.519	0.544	0.869

TABLE IV: Performance of different competing methods for binary classification.

Classifier	Micro				Macro			
	Precision	Recall	F1	ROC-AUC	Precision	Recall	F1	AUC
BotoM (DT)	0.770	0.770	0.770	0.760	0.760	0.760	0.760	0.760
SpamBot	0.680	0.680	0.680	0.660	0.727	0.665	0.646	0.660
FakeAcc (SVM)	0.696	0.696	0.696	0.701	0.704	0.701	0.693	0.701
NDSync	0.595	0.595	0.595	0.573	0.576	0.573	0.573	0.573
Decision Tree	0.859	0.859	0.859	0.633	0.640	0.632	0.622	0.632
K-NN	0.799	0.799	0.799	0.795	0.814	0.795	0.795	0.795
Logistic Regression	0.859	0.859	0.859	0.795	0.860	0.858	0.858	0.795
Naive Bayes	0.518	0.518	0.518	0.500	0.259	0.500	0.341	0.500
SVM	**0.873**	**0.873**	**0.873**	**0.873**	**0.874**	**0.873**	**0.873**	**0.873**
Random Forest	0.782	0.782	0.782	0.786	0.797	0.785	0.780	0.786
Bagging	0.695	0.695	0.695	0.698	0.709	0.698	0.684	0.698
Boosting	0.799	0.799	0.799	0.873	0.814	0.795	0.795	0.873

of Twitter users. It assumes that *a genuine user is less likely to post duplicate tweets as compared to a bot*. To compare our method with this baseline, we use their proposed set of features and their suggested classifier (Naive Bayes) to classify retweeters as 'Genuine', 'Bots', 'Promotional' and 'Normal'.

Baseline III: FakeAcc**:** [8] proposed a fake account detection model in Twitter based on minimum weighted feature set. It first measures gain ratio of 22 features and considers those features whose ratio crosses a certain threshold. Several classifiers are used, among which LR and SVM turned out to be the best for multi-class and binary classification respectively.

Baseline IV: NDSync**:** The closest baseline of our method is NDSync [10] which identifies multiple synchronous patterns across spam retweet threads. It extracts a set of features related to retweet threads (such as number of retweets, lifespan, response time etc.), projects retweet threads into a multi-dimensional feature space, and segments the feature space. It then estimates the suspiciousness score of each thread, and combines these scores for all the retweet threads associated with a user to obtain a user-level score. NDSync is an unsupervised method to classify retweeters as 'genuine' and 'fake'. We use this method for binary classification. However, for multi-class classification, we utilize the suspicious user score returned by NDSync as a feature for supervised classifiers.

B. Evaluation Setup

We combine the customers collected from blackmarket services (see Table II), yielding 743 customers (comprising of 86 bots, 275 promotional and 382 normal customers). We also sample 743 users randomly from a set of 1000 genuine users we collected, which makes the two classes balanced. The accuracy of each competing method is measured using the following metrics: Precision, Recall, F1-score, and Area under the ROC curve (AUC). Since multi-class classification in involved, we report all these metrics in both Micro and Macro settings (e.g., Micro-Precision, Macro-Precision etc.). We report the accuracy after averaging the result of 10-fold cross validation.

C. Results of Multi-class Classification

We design the first experimental setup as a four-class classification problem (genuine, bot, promotional, normal). Table III shows the results for the four-class classification. We obtain the best result of BotoM with Random Forest. Among four baselines, BotoM performs the best across all evaluation metrics. However, with our feature set, Logistic Regression achieves the maximum accuracy – it achieves 18% and 8.3% higher Micro-F1 and Micro AUC respectively, and 24.25% and 15.38% higher Macro-F1 and Macro AUC respectively w.r.t to the best baseline. The class-wise F1-score of Logistic Regression is as follows: 0.89 (genuine), 0.80 (bot), 0.58 (promotional), and 0.68 (normal).

D. Results of Binary Classification

One may only be interested to identify whether a retweeter is a customer or a genuine users, instead of a fine-grained classification of customers. Therefore, we conduct another set of experiments by combining all types of customers into a single class, and consider the problem as a binary classification problem. In Table IV, we notice that BotoM with Decision

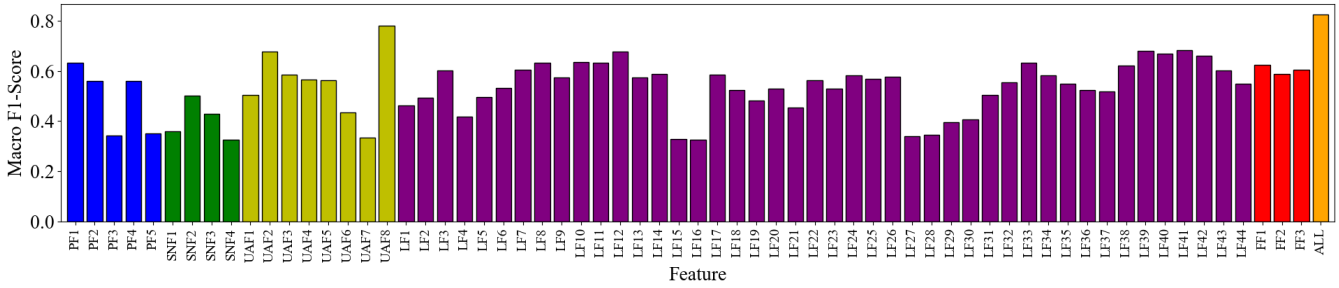

Fig. 4: Feature importance (accuracy of SVM for binary classification considering each feature in isolation). We also plot the accuracy with all features at the right-most bar of the figure (labeled as 'All').

Tree turns out to be the best baseline. However, this time SVM turns out to be the best model with a score of 0.873 for all metrics, which is followed by Logistic Regression. SVM beats the best baseline by 11.3% and 11.3% in terms of Micro-F1 and Micro AUC respectively, and by 10.3% and 11.3% in terms of Macro-F1 and Macro AUC respectively. The class-wise F1-score of Logistic Regression is as follows: 0.87 (genuine), 0.88 (customer).

E. Feature Importance

Figure 4 shows the importance of 64 features. Here we take each feature in isolation and run the best binary-classification model (SVM)[6]. The most important feature seems to be Bot-score (UAF_8) with which SVM achieves a Macro F1-score of 0.75. It also corroborates with the significantly high accuracy of our first baseline model (BotoM). The second and third ranked features are – maximum retweet likelihood in every Tuesday (LF_{39}) and in every Friday (LF_{42}) respectively. One possible reason why LF_{39} and LF_{42} have better discriminatory power is that these blackmarket services refresh their tweet database every 3-4 days (possibly every Tuesday and Friday). Thus, in order to keep their credit high, customers must keep on retweeting tweets every 3-4 days. However, as a whole, fluctuation features turn out to be the best (Macro F1=0.61), followed by user activity feature (Macro F1=0.554) and likelihood feature (Macro F1=0.55).

VI. Browser Extension: SCoRe

In the previous section, we have showed that state-of-the-art supervised models are capable of producing a significant accuracy with our proposed feature set to segregate blackmarket customers from genuine users. In order to help users deal with such fake retweeting activities, we attempt to build an extension, named SCoRe for chrome browser. It allows users to spot the blackmarket customers, which in turn provides a better way to understand the importance of a tweet.

Figure 5 shows a snapshot of how SCoRe facilitates users. The input is a link of the tweet whose retweeters the user wants to analyze. It first extracts all its retweeters, calculates the features for each retweeter, and then feeds it to the pre-trained SVM. Finally, the label (Genuine or Customer) of each

[6]The pattern of feature importance is same for the multi-class classification with Logistic Regression.

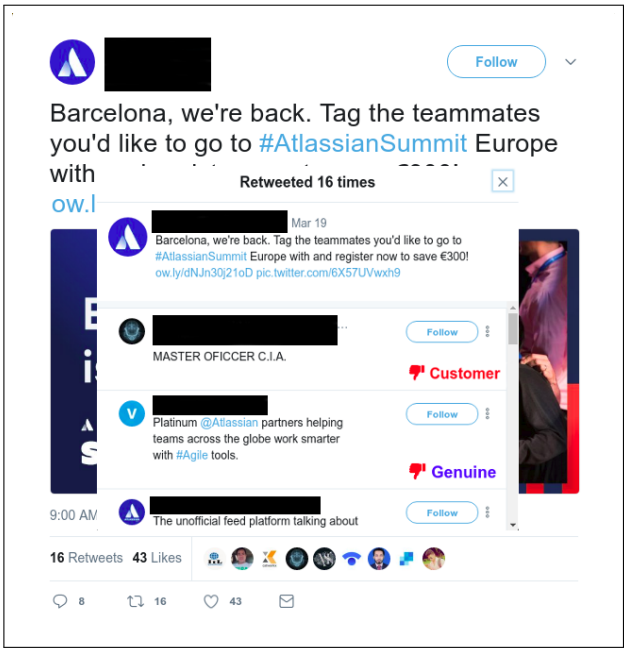

Fig. 5: Browser extension: SCoRe.

retweeter is displayed. It also allows users to provide feedback on each label. There is a 'thumbs down' symbol associated with each label, upon clicking of which the feedback will be forwarded to the back-end server. SCoRe has the capability to be trained incrementally with the feedback provided by the users. However, special care has been taken to make the learning process robust by ignoring spurious feedback. An attacker may want to pollute SCoRe by injecting wrong feedback. SCoRe handles these spurious feedback by first checking the confidence of the current model on labels associated with the feedback, and ignoring them if the confidence is more than a pre-selected threshold (currently set as 0.75). This makes SCoRe more robust under adversarial attacks.

User study: To analyse the performance of SCoRe, we conducted a user study with the help of 25 volunteers. First, we randomly assigned 20 tweet URLs to each volunteer. Volunteers visited each URL with the extension activated in their browser, clicked on the retweeter list (each retweeter was labeled as 'Genuine' or 'Customer' by SCoRe) and marked the wrong label by pressing 'thumbs down' button. SCoRe took 10 sec to label each retweeter. We recorded the responses of

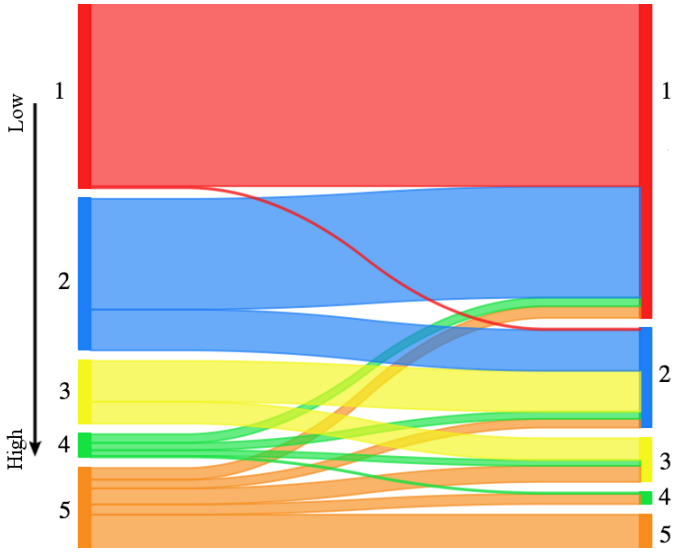

Fig. 6: **Alluvial diagram representing the flows of tweets between the bins (based on retweet count) before and after filtering customers.** The colored blocks correspond to different bins (the range of retweet count is divided equally into five bins). Left (right) blocks correspond to bins based on retweet rank before (after) the filtering. The size of the block indicates the number of tweets in that bin, and the shaded waves joining the regions represent flow of tweets between the bins, such that the width of the flow corresponds to the fraction of tweets.

the volunteers and measured the average accuracy. The results shows that SCoRe is highly accurate, achieving an average accuracy) of 85% (with standard deviation of 0.02).

VII. Implication of Collusive Retweeer Detection

Once we detect collusive retweeters, its immediate implication would be to re-rank the tweets based on modified retweet count (after filtering blackmarket customers). Figure 6 shows an alluvial diagram, indicating how tweets (500 tweets collected from the customers' timelines) change their ranking after such filtering. We show that top- and middle-ranked tweets are mostly affected by this filtering. We believe that this analysis opens the scope for a serious re-investigation of the existing metrics for ranking tweets/users.

VIII. Conclusion

In this paper, we studied the problem of understanding and detecting blackmarket-based collusive retweeters. The major contributions of this work are fourfold: **Dataset:** We collected a dataset of blackmarket collusive retweeters and annotated them. This, to our knowledge, is the first dataset of such kind. **Characterization:** We propose 64 novel features to characterize customers and differentiate them from normal users. **Classification:** State-of-the-art supervised methods performed significantly well to classify customers and genuine retweeters. **System design:** We developed SCoRe, a chrome extension that spots blackmarket retweeters in real-time.

Acknowledgement

The work was partially supported by Flipkart, India, Ramanujan Fellowship, and the Infosys Center for AI, IIITD.

References

[1] A. Aggarwal, S. Kumar, K. Bhargava, and P. Kumaraguru. The follower count fallacy: Detecting twitter users with manipulated follower count. In *ACM SAC*, 2018.

[2] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, page 12, 2010.

[3] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro. Aiding the detection of fake accounts in large scale social online services. In *USENIX*, pages 15–15, 2012.

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

[5] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In *WWW*, pages 273–274, 2016.

[6] C. De Micheli and A. Stroppa. Twitter and the underground market. In *11th Nexa Lunch Seminar*, volume 22, 2013.

[7] M. Eftekhar and N. Koudas. Some research opportunities on twitter advertising. *IEEE Data Eng. Bull.*, 36(3):77–82, 2013.

[8] A. ElAzab, A. M. Idrees, M. A. Mahmoud, and H. Hefny. Fake accounts detection in twitter based on minimum weighted feature. In *18th International Conference on Document Analysis and Recognition*, pages 1–6, 2016.

[9] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Comm. of the ACM*, 59(7):96–104, 2016.

[10] M. Giatsoglou, D. Chatzakou, N. Shah, A. Beutel, C. Faloutsos, and A. Vakali. Nd-sync: Detecting synchronized fraud activities. In *PAKDD*, pages 201–214, 2015.

[11] M. Giatsoglou, D. Chatzakou, N. Shah, C. Faloutsos, and A. Vakali. Retweeting activity on twitter: Signs of deception. In *PAKDD*, pages 122–134. Springer, 2015.

[12] S. Gupta, A. Khattar, A. Gogia, P. Kumaraguru, and T. Chakraborty. Collective classification of spam campaigners on twitter: A hierarchical meta-path based approach. In *WWW*, pages 529–538, 2018.

[13] X. Hu, J. Tang, Y. Zhang, and H. Liu. Social spammer detection in microblogging. In *IJCAI*, volume 13, pages 2633–2639, 2013.

[14] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang. Detecting suspicious following behavior in multimillion-node social networks. In *WWW*, pages 305–306. ACM, 2014.

[15] K. Lee, J. Mahmud, J. Chen, M. Zhou, and J. Nichols. Who will retweet this? detecting strangers from twitter to retweet information. *ACM TIST*, 6(3):31, 2015.

[16] Y. Liu, Y. Liu, M. Zhang, and S. Ma. Pay me and i'll follow you: Detection of crowdturfing following activities in microblog environment. In *IJCAI*, pages 3789–3796, 2016.

[17] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker. An analysis of underground forums. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 71–80. ACM, 2011.

[18] N. Shah, H. Lamba, A. Beutel, and C. Faloutsos. The many faces of link fraud. In *IEEE ICDM*, pages 1069–1074, 2017.

[19] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: growth and dynamics in twitter follower markets. In *ACM IMC*, pages 163–176, 2013.

[20] K. Thomas, C. Grier, D. Song, and V. Paxson. Suspended accounts in retrospect: an analysis of twitter spam. In *ACM IMC*, pages 243–258. ACM, 2011.

[21] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson. Trafficking fraudulent accounts: The role of the underground market in twitter spam and abuse. In *USENIX Security Symposium*, pages 195–210, 2013.

[22] A. H. Wang. Detecting spam bots in online social networking sites: a machine learning approach. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pages 335–342. Springer, 2010.

[23] A. H. Wang. Don't follow me: Spam detection in twitter. In *SECRYPT*, pages 1–10. IEEE, 2010.