Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Towards understanding bogus traffic service in online social networks[#]

Ping HE[1], Xuhong ZHANG[1], Changting LIN[2], Ting WANG[3], Shouling JI[†‡1]

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*

[2]*Binjiang Institute of Zhejiang University, Hangzhou 310027, China*

[3]*College of Information Sciences and Technology, Pennsylvania State University, University Park 17057-4846, USA*

[†]E-mail: sji@zju.edu.cn

Received Feb. 5, 2023; Revision accepted June 9, 2023; Crosschecked Feb. 20, 2024

Abstract: Critical functionality and huge influence of the hot trend/topic page (HTP) in microblogging sites have driven the creation of a new kind of underground service called the bogus traffic service (BTS). BTS provides a kind of illegal service which hijacks the HTP by pushing the controlled topics into it for malicious customers with the goal of guiding public opinions. To hijack HTP, the agents of BTS maintain an army of black-market accounts called bogus traffic accounts (BTAs) and control BTAs to generate a burst of fake traffic by massively retweeting the tweets containing the customer desired topic (hashtag). Although this service has been extensively exploited by malicious customers, little has been done to understand it. In this paper, we conduct a systematic measurement study of the BTS. We first investigate and collect 125 BTS agents from a variety of sources and set up a honey pot account to capture BTAs from these agents. We then build a BTA detector that detects 162 218 BTAs from Weibo, the largest Chinese microblogging site, with a precision of 94.5%. We further use them as a bridge to uncover 296 916 topics that might be involved in bogus traffic. Finally, we uncover the operating mechanism from the perspectives of the attack cycle and the attack entity. The highlights of our findings include the temporal attack patterns and intelligent evasion tactics of the BTAs. These findings bring BTS into the spotlight. Our work will help in understanding and ultimately eliminating this threat.

Key words: Online social networks; Measurement; Bogus traffic; Black market
https://doi.org/10.1631/FITEE.2300068 **CLC number:** TP39

1 Introduction

The hot trend/topic page (HTP), e.g., trends in Twitter and hot search in Weibo, is an essential mechanism for the functionality and underlying business model of microblogging sites. These pages aggregate mainly the popular topics that have spikes in the platform's current traffic volume. Numerous platform users check the HTP to obtain the current hot news and information of interest. However, once disinformation is illegally pushed into the HTP, it will be further amplified (Elmas et al., 2021), and many users might be misled by the disinformation. Therefore, the integrity and authenticity of the HTP are integral to microblogging sites. Unfortunately, the integrity and authenticity of the HTP are severely damaged by an emerging underground service that hijacks topics in the HTP.

In this paper, we refer to this kind of underground service as a bogus traffic service (BTS). BTS agents rely on controlling and maintaining an army of

[‡] Corresponding author

[#] Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2300068) contains supplementary materials, which are available to authorized users

ORCID: Ping HE, https://orcid.org/0009-0004-9911-7897; Xuhong ZHANG, https://orcid.org/0000-0002-8571-9780; Changting LIN, https://orcid.org/0000-0002-8918-6299; Ting WANG, https://orcid.org/0000-0003-4927-5833; Shouling JI, https://orcid.org/0000-0003-4268-372X accounts that we call bogus traffic accounts (BTAs) to orchestrate the entire attack. BTAs tweet and retweet the tweets containing the target topic, generating a vast amount of traffic related to the topic in a short time. The burst of traffic fools the underlying HTP system into including the topic. BTS agents can be easily found from various channels such as websites, e-commerce platforms, and even in online social networks (OSNs). The convenience and low cost of BTSs lower the entry barriers for manipulating public opinion and twisting facts, e.g., making it possible to guide public opinion and influence political elections (Just et al., 2012).

Previous research has studied Twitter spam markets (Thomas et al., 2013), Twitter follower markets (Stringhini et al., 2013), Facebook like farms (de Cristofaro et al., 2014), and collusive retweeting markets (Dutta and Chakraborty, 2020). The goal of these black markets is mainly to increase the numbers of likes and followers, whereas the BTS aims at pushing a fake topic into the HTP. Moreover, BTS requires agents to adopt more complex operating mechanisms to control the BTAs to conduct attacks and evade detection. So far, to the best of our knowledge, no prior work has been dedicated to studying BTSs.

To bridge the gap, in this paper, we conduct a large-scale comprehensive measurement study of BTSs, which answers mainly three questions: What is the BTS marketplace? How can we efficiently identify the bogus traffic? What is the BTS operating mechanism?

To answer the first question, we conduct an integrated and detailed investigation of BTS agents. We first explore three different approaches to expose as many agents as possible, including querying search engines, searching on e-commerce platforms, and analyzing the promotional profile portraits from social networks. In total, we collect 125 agents with their prices and contact information (Section 4).

To answer the second question, we build a machine learning classifier to detect BTAs as a bridge to identify the bogus traffic. The motivation for this indirect approach is that the bogus traffic content, such as fake retweets, is not significantly different from the tweets generated by normal users, although there are still some linguistic characteristics that can uniquely identify BTAs (Section 5).

To answer the last question, we deploy our classifier on our large-scale dataset containing 523 323 users and detect 162 218 BTAs in it with a precision of 94.5%. Then we set up some rules to filter out the tweets related to the BTS. Based on these BTAs and tweets, we uncover the operating mechanism of BTS from the perspectives of the attack cycle and the attack entity. Highlights of our findings include the temporal attack patterns and the intelligent evasion tactics of BTAs (Section 6).

Our main contributions are summarized as follows:

1. We perform a 10-month study to build a ground-truth dataset containing 5042 BTAs and 6652 benign users, and a large-scale dataset containing 523 323 accounts and 80 182 376 tweets. We will share these datasets to facilitate future research.

2. By analyzing the linguistic differences between normal tweets and evasive tweets, we find a unique text-based feature to distinguish BTAs from normal users. Combined with other profile-based features, we propose an effective BTA detection method that can be integrated into current monitoring systems.

3. We find a new kind of temporal attack pattern. BTAs will generate a large number of fake retweets in the beginning and at the end, but few retweets in the middle of an attack cycle, whereas existing malicious identities, e.g., fraud likers (de Cristofaro et al., 2014), collusive retweeters (Dutta and Chakraborty, 2020), CrowdTurfers (Song et al., 2015), and black-market accounts (Song et al., 2015), either perform malicious activities steadily or concentrate only on one period of time. The fundamental reason for this new temporal attack behavior is that the fake traffic generated by the BTS could also attract real traffic.

4. We observe a series of new BTA evasion tactics. For example, they have limited connections with other users and do not explicitly perform malicious actions against their surrounding network, such as spreading malicious uniform resource locators (URLs) (Thomas et al., 2014). In addition, their evasive tweets are very human-like, because they are mainly from a large corpus crawled from famous quote websites and news websites. Even in their fake retweets, the additional comments added by them are highly relevant to the original tweets.

2 Related works

In this section, we categorize related works into three categories corresponding to the focus of our research questions: marketplaces, detection, and operating mechanisms.

2.1 Black marketplace discovery

Understanding and measuring the black market has been an active research area for a long time. van Wegberg et al. (2018) evaluated the commoditization of cybercrime and discovered that the cash-out service feature involves the most listings and generates the largest revenue. Booij et al. (2021) investigated the vendor careers of the dark net market and found that their career trajectories were heavily unbalanced in terms of longevity and success. Cuevas et al. (2022) conducted a comprehensive analysis of the Hansa market and found its total market revenue to be 50 million US dollars.

In the OSN domain, Dutta and Chakraborty (2020) described the characteristics of collusive retweeters. They discovered that black market services have two types: premium and freemium. Torres-Lugo et al. (2022) analyzed the US hyperpartisan train network and found evidence of activity by inauthentic automated accounts and abnormal content deletion. None of them are dedicated to BTS which aims to push topics to the HTP. In addition, BTS is a more advanced malicious service that employs various intelligent evasion tactics to conduct attacks on behalf of customers.

2.2 Fraud account detection

Existing fraud account detection methods for OSNs can be divided roughly into three categories: feature-based approaches, graph-based approaches, and aggregate behavior based approaches.

1. Feature-based approaches. Feature-based approaches (Song et al., 2015; Beskow and Carley, 2019, 2020) model fraud account detection as a binary classification problem and adopt machine learning techniques. Specifically, they first encode a user's behavioral patterns and profiles in features. Then they leverage supervised machine learning techniques or unsupervised machine learning techniques to detect sybils. For instance, CrowdTarget (Song et al., 2015) detects crowdturfing tweets based on features from retweet time distribution.

2. Graph-based approaches. Graph-based methods (Ali Alhosseini et al., 2019; Feng et al., 2021, 2022, 2023) view accounts as nodes and social links between accounts as edges. They detect fraud accounts by leveraging graph algorithms, e.g., graph convolution neural networks (Ali Alhosseini et al., 2019). These works often hold the assumption that in a social graph, there exist a limited number of attack edges between benign users and fraud accounts (Feng et al., 2022).

3. Aggregate behavior based approaches. Aggregate behavior based approaches (Cao et al., 2014; Zheng et al., 2017; Yuan et al., 2019) focus on uncovering a group of users by identifying the synchronized group activities. For example, Ianus in Yuan et al. (2019) extracts common registration patterns, e.g., using the same Internet protocols (IPs) at registration time.

Unfortunately, fraud accounts are becoming increasingly advanced and hard to identify (Freitas et al., 2015; Cresci et al., 2017, 2019; Weerasinghe et al., 2020). They perform complex behaviors, e.g., copying personal pictures (Woolley, 2016), and conduct more difficult attacks, e.g., trend manipulation (Zhang et al., 2017; Guo et al., 2018). These complex behaviors bring new patterns to fraud accounts, which makes these detection methods ineffective.

2.3 Operating mechanism disclosure

Zhang et al. (2017) discussed the Twitter trend manipulation problem and revealed that popularity, coverage, potential coverage, and reputation are the key factors in the Twitter trend. Guo et al. (2018) used a specific Weibo hot topic to demonstrate that spammers can manipulate public sentiment. Jakesch et al. (2021) analyzed manipulated Twitter trends in the Indian general election and showed that the campaigns produced hundreds of nationwide Twitter trends throughout the election. These works analyzed only the operating mechanism on a limited dataset (e.g., 2000 accounts), and they did not analyze the patterns from the malicious account side (e.g., temporal activities and evasiveness). We conduct a more comprehensive study on a large-scale dataset with more than 5×10^5 accounts and more than 8×10^7 tweets.

The work by Elmas et al. (2021) is perhaps the most related to ours. Elmas et al. (2021) presented an ephemeral astroturfing attack on Twitter trends. However, the astrobots in their work are not as intelligent and evasive as our BTAs. For example, the astrobots post ungrammatical tweets by randomly concatenating words and phrases. To evade detection, they simply delete fake tweets in a few minutes.

3 Background

3.1 Microblogging sites

Microblogging sites are particularly popular nowadays. Users can post their tweets, view others' tweets, and like or retweet others' tweets. To improve user experience, microblogging sites aggregate the topics that are being intensely discussed on the platform, such as trends in Twitter and hot search in Weibo. The inclusion of a topic in an HTP is generally determined by a combination of its traffic volume and the time it takes to create the volume. For example, Weibo sorts topics based on the increment of topic reading per hour and the amount of user participation per hour, and then lists the top 50 as hot search topics. The HTP, as one of the top-level entries of these platforms, is heavily viewed by users to get to know the hot news. Thus, if a malicious or fake topic appears in the HTP, an enormous number of users will be affected.

3.2 Threat model

The BTS threat model is shown in Fig. 1. It consists mainly of two parts: the physical world and the OSN world. This study focuses on attacks in the OSN world. We define the attack process as the following: BTAs (attacker/attack entity) massively tweet and retweet with the target topics (hashtags) until the site's HTP (victim) includes the target topics. Meanwhile, the BTAs adopt complex tactics to impersonate benign accounts to evade detection.

Specifically, malicious customers want to push a target topic into the HTP. For instance, the target topic may be related to a superstar. The malicious customers achieve their goal by purchasing the BTS from the agents. These agents can be discovered through channels, including search engines and e-commerce platforms, and within the OSN itself.

These BTS agents control and maintain an army of BTAs, which execute the concerted operation. The BTAs adopt various strategies to mimic genuine accounts within the OSN, enhancing their semblance of normality. During the execution of this attack, the BTAs push the target topic into the HTP by generating an immense volume of traffic related to the topic within a condensed time frame. Once the selected topic has gained sufficient traffic to appear in the HTP, the benign accounts on the platform can view this topic within the HTP.

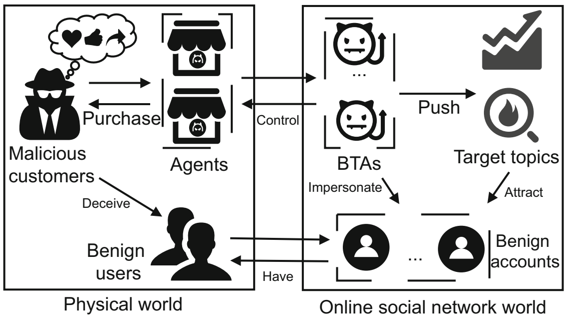

Fig. 1 Threat model of our research consisting of malicious customers, agents, benign users, BTAs, target topics, and benign accounts

If these unsuspecting benign users opt to delve into the specifics of the topic, they are confronted with the content desired by the malicious actors. It should be noted that the quality and authenticity of the content are not guaranteed. The attack serves to mislead the HTP, thereby suppressing organically popular topics. Consequently, benign users are unwittingly deceived by malicious customers.

4 Marketplace of BTS

In this section, we first describe the procedure for finding the BTS agents, in which we observe some interesting promotional strategies that may be designed to bypass the current censoring systems. Then we investigate their preferred communication channels and analyze the possible reasons from three perspectives.

4.1 Finding agents of BTS

The first step in our investigation is to find BTS agents. Because this kind of service is an underground service, it is difficult to directly find those agents. We try to find them using three sources: search engines, e-commerce platforms, and Weibo.

For search engines, we use some bogus traffic-related keywords (e.g., buy traffic, HTP service, and HTP business) to search potential BTS websites. For each search result, we manually check if the website offeres BTS. Finally, to the best of our effort, we find 11 agents from these websites.

For e-commerce platforms, we use the same strategy to search bogus traffic related keywords in Taobao, Jingdong, and Pinduoduo, because they are the mainstream e-commerce platforms in China. For instance, Taobao has 726 million consumers in China (Alibaba Inc., 2020); Jingdong has more than 270 000 third-party merchants (JD Inc., 2020); Pinduoduo has more than 300 million active buyers and more than 1 million merchants (PDD Inc., 2020). Finally, to the best of our effort, we find 12 agents on these three e-commerce platforms. It is worth mentioning that all the agents we find on these e-commerce platforms post little text information about what they could do. Almost all the text information is simply a message "Please contact for details." Instead, they put all the BTS details into images. We believe that it is because e-commerce platforms have better content censoring on text descriptions and pay less attention to the descriptions in images.

For Weibo, we randomly collect 275 719 user profile portraits in Weibo and manually identify the text information in them, from which we obtain 2310 photos containing BTS-related promotional information. We also find an interesting promotional strategy of the agents, i.e., hiding the promotion information in user profile portraits. Fig. 2 shows some promotional profile portraits. In these photos, the promotion information is hidden in the original photo in a watermark form. The promotion information usually has two core parts, including the product descriptions and contact methods. In addition, the texts are deformed, e.g., spinning and bending, to make them hard to identify. We manually contact all agents who left the contact information in these photos. In total, we discover 102 BTS agents. We also evaluate the effectiveness of existing optical character recognition (OCR) techniques in extracting the text information from these portraits. Specifically, we test 100 photos by using the state-of-the-art OCR technique offered by machine learning as a service (MlaaS) from Baidu Cloud, Ali Cloud, and Huawei Cloud. Test results show that they can recognize only about 60% of the words on average, which means that so far state-of-the-art OCR algorithms cannot solve this problem well.

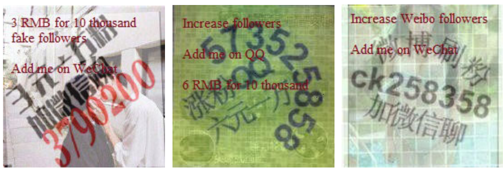

Fig. 2 Promotional profile portraits

Promotional information is hidden in the picture in the form of a watermark. The words in red are the translations of the Chinese characters in the watermark. References to color refer to the online version of this figure

By chatting with these agents, we find that they also offer other malicious services, such as fake following, fake liking, and fake search volume. We observe that the agents use the same group of accounts to perform these tasks. For instance, we find that the fake followers in the fake following service also post fake tweets for BTS.

4.2 Communication channels

BTS agents need a way to communicate with their customers. To find out how we could contact them, we manually analyze the contact methods shown in the promotional profile portraits. We find that the BTS agents use only instant messaging applications (IM Apps) as their communication channels, especially QQ and WeChat, which are the most widely used applications in China. Specifically, 71.6% of them used QQ and 58.8% of them used WeChat. Surprisingly, no black forum is used as a communication channel. We analyze the reasons why IM Apps such as QQ and WeChat are adopted as the main communication channel from the following three perspectives: popularity, security, and convenience. Detailed analysis can be found in the supplementary materials.

5 Identifying bogus traffic

Directly identifying bogus traffic is a challenging task, because its contents, e.g., fake retweets, are almost the same as those from normal users. Therefore, we propose to detect the BTAs first as a bridge to identify the bogus traffic. Although the BTAs adopt evasive tactics, we find that there are still some linguistic characteristics that can uniquely identify them. Specifically, we take a further look at their evasive tweets (examples are given in the supplementary materials), which are not involved in any bogus traffic (no hashtag in the tweet) and are used to make the BTAs look like human users. We find that these evasive tweets are linguistically different from the tweets generated by normal people.

Fig. 3 shows our overall methodology for detecting BTAs. We first present our infiltration framework to gather the initial seed BTA dataset. Then, based on the analysis of the seed dataset, we propose a novel detection algorithm that consists mainly of a BERT-based feature extractor and an XGBoost classifier. The BERT-based feature extractor is used to learn the linguistic differences between evasive tweets and normal tweets. Finally, the XGBoost classifier combines the linguistic feature and other profile features to make the final decision.

5.1 Infiltration framework

Recall that we have found 125 BTS agents (Section 4). To build the ground-truth BTA dataset, we choose to purchase BTS from these agents for research purposes. We discuss the potential ethical problems in Section 7.3.

Table 1 shows our purchase solution. We could not purchase BTS from any discovered agents because of the financial cost and ethical problems. Considering the financial cost, we purchase BTS from seven randomly selected agents in our list. Instead of directly purchasing BTS, we purchase an alternative malicious service, i.e., a fake following service, which the BTS agents also offer. This is because we aim at obtaining the maximum number of BTAs. For the fake following service, we could easily obtain the accounts that follow us. However, in other alternative services, such as fake searching, fake liking, or BTS, we could not determine who performes the task. Another reason is that the fake following service is the least harmful service compared to other alternative services. This service is cheaper than other services, which creates the least revenue for underground markets. In our experiments, we set up a honeypot Weibo account to collect the purchased accounts. To further reduce the potential impact on Weibo, we set an announcement stating that the account is used for bogus traffic research: please do not trust the messages sent by all the fans of this account. We also leave our contact method in this announcement and report it to Weibo. Another interesting finding in our purchase is that the agents are dishonest. As shown in Table 1, we can see that the number of actually delivered BTAs is much smaller than the number of accounts ordered. As a result, we successfully obtain 5042 BTAs out of the 7450 ordered BTAs.

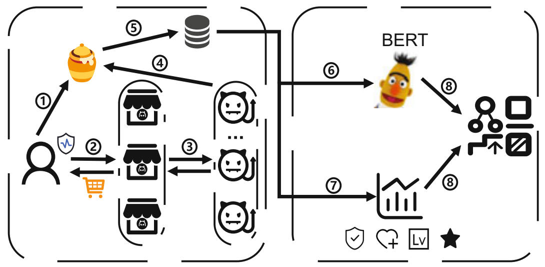

Fig. 3 Overview of our detection methodology

①: setting up a honeypot account to capture BTAs. ②: purchasing BTAs from agents. ③: agents manipulate BTAs to conduct an attack. ④: BTAs' attack is captured by the honeypot account. ⑤: building the ground-truth dataset. ⑥: using BERT to extract semantic features. ⑦: analyzing profile-based features. ⑧: XGBoost classifier

In addition to the seed ground-truth BTAs purchased from the agents, we collect the fans of these seed BTAs as a candidate account dataset for our measurement study. Our intuition is that the fans of BTAs are more likely to be BTAs. We believe that if we could learn a detector from the seed BTAs, we could use it to identify more BTAs from this candidate dataset to enlarge the dataset used in our measurement study. In total, we obtain 523 323 fans of the seed BTAs to form the candidate account dataset. In the process of collecting benign users in Weibo, we observe a phenomenon which could help us reduce the potential false positives. Specifically, we find that benign users will communicate with others in Weibo. Thus, we collect only the users who comment on the hot tweets and also communicate with others in the comments. In total, we collect 6652 benign users. After obtaining all the seed BTAs, candidate accounts, and benign users, we use web crawlers to collect their information in Weibo, which consists of two parts: profile information and the posted tweets. The crawling period for the seed BTAs and benign users was from August 2019 to December 2019, and the crawling period for the candidate accounts was from January 2020 to May 2020. Table 2 summarizes the information for the datasets obtained in our infiltration process.

5.2 Detecting BTAs

After analyzing the behaviors of the seed BTAs, we find that they adopt tactics to evade detection;

Table 1 Purchase solution for BTAs

Agent	Source	Number of accounts ordered	Number of accounts obtained
GYCM87	Head portraits	50	27
KunMaiFen	Head portraits	1500	1004
June Flower Outdoor Store	E-commerce	700	468
Yetian Outdoor Store	E-commerce	700	489
Yesheng Outdoor Store	E-commerce	500	473
fsj3.com	Website	2000	1378
www.niufenba.com	Website	2000	1203
Total		7450	5042

Table 2 Datasets gathered in our infiltration process

Dataset	Type	Number of accounts	Time
Ground-truth dataset	BTAs	5042	Aug. 2019–Dec. 2019
	Benign users	6652	Aug. 2019–Dec. 2019
Candidate dataset	Fans of BTAs	523 323	Jan. 2020–May 2020

e.g., they will post many tweets such as famous quotes which are human-generated tweets. Moreover, the tweets related to bogus traffic are only a small portion of all the tweets they post.

In particular, their malicious behaviors, i.e., fake retweets, do not exhibit aggregate behavioral patterns, which are considered as the core features in previous sybil detection methods (Cao et al., 2014; Thomas et al., 2014; Zheng et al., 2017). To verify this finding, we perform a preliminary study on the seed BTAs by computing the pairwise behavioral similarity of BTAs from the same agent. Specifically, we denote one malicious behavior of the BTAs as a tuple (U, J, T), where U, J, and T represent the account identity, target topic in the tweet, and timestamp of the tweet, respectively. For accounts u and v, we derive the behavior sets $B(u)$ and $B(v)$ associated with u and v as

$$B(u)=\{(U,J_1,T_1),(U,J_2,T_2),\cdots,(U,J_m,T_m)\}, \quad (1)$$

$$B(v)=\{(V,J'_1,T'_1),(V,J'_2,T'_2),\cdots,(V,J'_n,T'_n)\}. \quad (2)$$

For pairwise accounts u and v, and for a given k, $(U, J_k, T_k) \in B(u)$, we specify that if $P_u(k) = 1$ exists, $(V, J'_l, T'_l) \in B(v)$, such that the following two proprieties are true:

(1) The two target topics are the same, $J_k = J'_l$;

(2) The two behaviors occur within a fixed time slot $\Delta T, |T_k - T'_l| \le \Delta T$.

$P_u(k) = 0$ otherwise. Finally, we compute the similarity between pairwise accounts u and v as follows (Zheng et al., 2017):

$$\mathrm{Sim}(u,v) = \frac{\sum_{k=1}^{m} P_u(k) + \sum_{l=1}^{n} P_v(l)}{|B(u)| + |B(v)|}.$$

The similarity computes the percentage of possible collaborative tweets in the tweets of two accounts. If two accounts are highly similar, then we say that they have a stronger aggregate behavioral pattern. Then, we set $\Delta T = 24$ h (Zheng et al., 2017) and compute the cumulative distribution function (CDF) of the pairwise BTA similarities for each agent in our seed BTA dataset. As shown in Fig. 4, overall, about 70% of the pairwise similarities are zero. This indicates that the BTAs do not have aggregate behavioral patterns.

To defend against this high evasiveness, we go one step further by looking at the contents of the evasive tweets. Then we discover an untapped strong indicator—the linguistic differences of the evasive tweets posted by the BTAs. The motivation for this choice is our finding that these evasive tweets are not original; i.e., they are sampled from a specific corpus maintained by the BTS agents. The corpus includes many quotations expressed in ancient Chinese, e.g., “Sharpening makes a mighty sword, and cold makes a blooming wintersweet.” These tweets have differences in word choice and sentence structure compared with the tweets generated by normal people, e.g., “The grind of stone sharpens the blade, the bite of frost sweetens the wintersweet.” Thus, they should have linguistic differences with the ones generated by normal people.

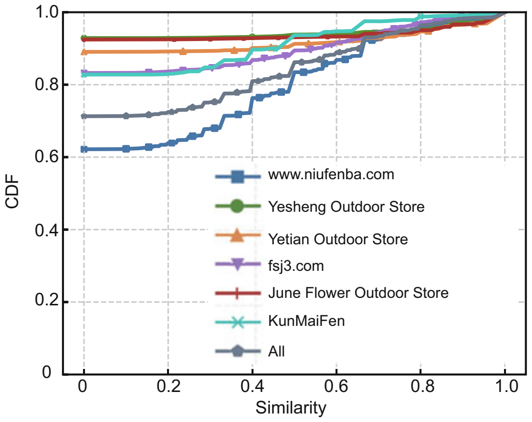

Fig. 4 Cumulative distribution function (CDF) of pairwise behavioral similarity of BTAs from different agents in our ground-truth dataset (A higher similarity indicates a stronger behavioral aggregate pattern)

Therefore, we take the linguistic differences of the evasive tweets into consideration when detecting the BTAs. Specifically, we customize BERT (Devlin et al., 2018) to build a classifier to predict if a tweet is an evasive tweet. This is because the BERT model is one of state-of-the-art natural language processing (NLP) models (Liu et al., 2023). BERT has set new standards for precision across 11 typical NLP tasks, consistently outstripping other top-tier NLP models in terms of performance, and notably being the first to surpass human benchmarks. It is for these reasons that we propose BERT as the optimal solution for identifying evasive tweets with a high accuracy. We also try to use another NLP model called word2vec (Mikolov et al., 2013; Le and Mikolov, 2014) with a logistic regression classifier as the comparison method to help us accomplish the evasive tweet detection task. However, the results indicate that this approach yields suboptimal performance in contrast to our primary BERT-based model.

In the task of evasive tweet detection, our objective is to train the customized BERT model to accurately differentiate between evasive and normal tweets. To accomplish this, we append a multilayer perceptron (MLP) model to the original BERT model. This MLP component is primarily responsible for executing the classification task, while a cross-entropy loss function serves as the optimization method for the MLP model. Our training dataset is composed of two types of samples. Evasive tweets are designated as positive samples and normal tweets as negative samples. As the training process unfolds, it is anticipated that the model will increasingly learn to discern evasive tweets from their normal counterparts.

We follow the intuition that if an account has a high percentage of evasive tweets, it is highly likely to be a BTA. Therefore, we estimate the percentage of evasive tweets generated by BTAs in an account. To further improve the precision and recall of our BTA classifier, we combine it with five other profile-based features, as described in the following to build an XGBoost (Chen and Guestrin, 2016) detector to predict whether an account is a BTA or not. The five profile-based features are introduced in the supplementary materials. The rationale behind our choice of the XGBoost model rests on its commendable explanatory capacity. In contrast to neural networks, XGBoost is inherently interpretable, offering insights into the importance of individual features. Consequently, the platform can use these explanations to provide justifications and send notifications to the accounts that have been detected. Additionally, our method is evaluated in comparison with alternative machine learning classifiers such as support vector machines, random forests, and MLP.

With regard to the BTA detection task, the training objective is to enable the classification model to differentiate between standard users and BTAs. The training dataset encompasses features derived from both standard users and BTAs. Throughout the training process, it is envisaged that the model will progressively learn to discern between normal users and BTAs.

5.3 Evaluation

1. Hardware. We ran all the experiments on a server with Intel Xeon CPU E5-2650, four NVIDIA GeForce GTX 1080 Ti GPUs, and 256 GB RAM.

2. Evasive tweet detection. The primary feature that our BTA detection model relies on is the percentage of evasive tweets in an account. Thus, we first evaluated the performance of the BERT-based classifier. We randomly sampled 3047 BTAs from the 5042 seed BTAs. Then we selected all the tweets without hashtags from these accounts as the positive samples for this classifier. In total, we obtained 34 358 positive samples. As for the negative samples, we randomly sampled 3047 users from the 6652 benign users. Similarly, we selected all the tweets

without hashtags from them as negative samples. In total, we obtained 15 096 negative samples. Then, we divided these tweets into the training set, validation set, and test set according to the ratio of 6:3:1. To prevent overfitting, we fine-tuned only this classifier on the BERT-base-Chinese model (HuggingFace, 2022), which has been pre-trained for Chinese by the HuggingFace team for one epoch. Our evasive tweet classifier achieved 92.3% precision on the validation set and 92.8% precision on the test set. We also evaluated the word2vec model on the same dataset. Specifically, we used the word2vec model to capture the word embedding of every word in the sentence and computed the average of the word embeddings as the sentence embedding. Then, we trained a logistic regression classifier using sentence embedding to classify evasive tweets and normal tweets. As a result, the word2vec model with a logistic regression classifier achieved a precision of only 85.3% on the test set, which is lower than that of our customized BERT model (92.8%).

Moreover, we visualized the embedding of the tweets from the BTAs and benign users. Fig. 5 is the visualization of the last layer output of the evasive tweet classifier for 50 random positive samples and 50 random negative samples from our training set. Fig. 5 shows that the positive samples and negative samples have a clear separation, which indicates that our model has learned the linguistic differences between the evasive tweets and normal tweets.

Finally, we analyzed the effectiveness of using the percentage of evasive tweets in an account as a feature to detect BTAs. Specifically, we randomly sampled 500 BTAs from the remaining 1995 BTAs and 500 benign users from the remaining 3505 benign users. Then we computed the percentage of evasive tweets for each of them and displayed their distributions in Fig. 6. From Fig. 6, we can see that the distributions of this feature for the BTAs and benign users are significantly different, which indicates that this feature could be a strong signal in detecting BTAs.

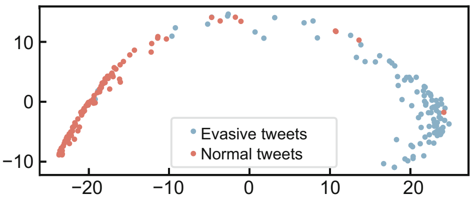

Fig. 5 Evasive tweet embedding visualization

References to color refer to the online version of this figure

3. BTA detection. To train and test the XGBoost (Chen and Guestrin, 2016) model, we used the remaining 1995 BTAs and an equal number of benign users from the remaining 3505 benign users. Then we divided the dataset into a training set and a test set according to the ratio of 7:3. The results showed that our model achieved 97.2% precision and 95.9% recall on the test set. Furthermore, we checked the feature importance among the six features with the results shown in Fig. 7. From Fig. 7, we can see that the percentage of evasive tweets is the core feature to distinguish BTAs from benign users. To further prove the importance of the text-based feature, we trained and tested another XGBoost model that used only the selected five features in the same dataset as the baseline model. It achieved only a precision of 91.3%.

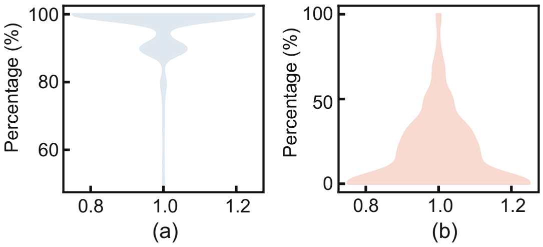

Fig. 6 Distribution of the evasive tweets: (a) BTAs; benign users

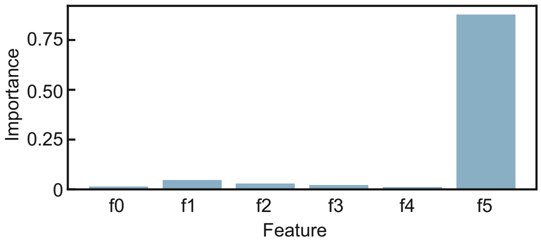

Fig. 7 Feature importance of our BTA detector

f0 is the authentication, f1 is the account level, f2 is the number of followings, f3 is the number of followers, f4 is the number of tweets, and f5 is the percentage of evasive tweets

In addition to our primary method, we examined other similar algorithms as comparative classifiers for BTA detection, using the same training dataset. Specifically, we deployed the word2vec model as a linguistic feature extractor in conjunction with various other machine learning algorithms to identify BTAs within the dataset. The outcomes of these comparisons are illustrated in Table 3. As

Table 3 Performance of the BTA detection task on different methods

NLP model	Classifier	Precision	Recall	F1
BERT	XGBoost	0.972	0.959	**0.965**
	SVM	0.950	0.957	0.954
	RF	0.971	0.945	0.960
	MLP	0.958	0.957	0.957
word2vec	XGBoost	0.948	0.973	0.960
	SVM	0.868	0.911	0.889
	RF	0.961	0.937	0.949
	MLP	0.877	0.911	0.894

Best result is in bold

shown in Table 3, it is evident that the best performance is achieved through the combination of the BERT model and the XGBoost classifier. Although it is noteworthy that other methods also exhibit satisfactory performance on the task of BTA detection, we argue that this efficacy can be largely attributed to the role of linguistic features, a key element in our study's findings.

6 Operating mechanism of BTS

6.1 Landscape

After deploying our method on the large-scale candidate dataset in Table 2, we identified 162 218 BTAs. To test the precision of our method on this candidate dataset, we randomly sampled 1000 BTAs from the identified BTAs and manually verified each of them according to our verification rules. The rules are described in the supplementary materials. The estimate precision was 94.5%. Then we collected all the tweets containing hashtags from these detected BTAs as the bogus traffic-related tweets. In total, we obtained 3 591 869 tweets containing 296 916 unique topics. In our study, we considered each topic as an attack. Finally, based on these detected BTAs and their associated bogus traffic-related tweets, we conducted a measurement study of the BTS operating mechanism. In particular, we first examined the customers of the BTS. Then, we randomly selected 1000 topics in the collected bogus traffic and found that the customers of BTS were from a variety of professions including luxury sale, entertainment, sports, and tourism. In the following, we will dissect the operating mechanism from the perspective of the threat model defined in Section 3, including the attack cycle and attack entity.

6.2 Attack cycle

To better understand the attack from the time dimension, we studied the duration of each attack by computing the time difference between the first tweet and last tweet associated with each attack. Then we plotted the CDF of the attack duration in Fig. 8a. We can see that about 90% of the attacks finished in 10 days. This short duration property reflects the ad-hoc nature of this kind of attack. The malicious customers of BTS only want to catch public attention, and once this goal is reached, they will stop the attack. For example, once the desired topic is pushed into the platform's HTP, the attack will cease. On the other hand, the short-term attack makes the appearance and disappearance of the topic more natural, thereby increasing the credibility of the topic itself.

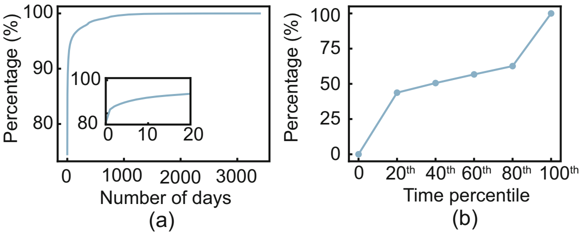

Fig. 8 Attack cycle measurement: (a) CDF of attack cycle duration; (b) CDF of bogus traffic volume within an attack cycle

To perform a more detailed analysis of the time pattern of this attack, we estimated the attack intensity within each attack cycle. The attack intensity can be reflected by the bogus traffic volume, e.g., the number of tweets containing the target topic, within a certain time. We accumulated the average bogus traffic volume ratio at different time percentiles of the attack cycle for all the attacks. For example, the 20th percentile of a 5-day attack cycle represents the first day of the attack. To reduce the impact of noise, we selected only the attacks that lasted >5 days. Fig. 8b shows this distribution. We can see that most of the bogus traffic concentrates around the beginning and end of an attack cycle. This means that even within the short attack cycle, the attackers used smart strategies to optimize their resource. The underlying reason for this intermittent attack intensity could be that the initial round of an attack will attract a large amount of real traffic from normal users to spread the target topic. Thus, the attacker needs to generate less bogus traffic to keep the popularity. Once the popularity of the target topic cools down, the attacker will need to invest more to maintain the popularity. Moreover, we use the topics related to superstars as a case to show the patterns of the attack cycle. Details are given in the supplementary materials.

6.3 Attack entity

BTAs are the main entities that generate bogus traffic. Analyzing them is critical for a deeper understanding of the operating mechanism of BTS. Specifically, we present our study of their influence in the network, life cycles, and evasion tactics in this subsection.

6.3.1 Influence analysis

To measure the influence of BTAs in social networks, we plotted the CDF of the numbers of likes, comments, and retweets per tweet associated with an account. Intuitively, more likes, comments, and retweets indicate a more influential user. As shown in Fig. 9, on these three metrics, the values of BTAs are significantly lower than those of benign users, which indicates that the influence of the BTAs is much lower than that of benign users. For example, in Fig. 9a, nearly all of the BTAs have zero likes per tweet, whereas >60% of benign users have more than 1 like per tweet. This indicates that no one will pay attention to the tweets posted by BTAs. We can draw similar conclusions from the two other metrics. To see whether BTAs will spread malicious URLs to others as reviewed in previous works (Yang et al., 2013; Thomas et al., 2014; Song et al., 2015), we constructed a regular expression to match the URLs in their tweets. However, we did not find any URL. This indicates that BTAs do not use social connections to spread malicious URLs. From the two phenomena discussed above, we conclude that BTAs are harmless to social connections, because they do not leverage social connections to benign users to form the attack edges as revealed by previous works (Yu et al., 2006, 2010; Alvisi et al., 2013; Boshmaf et al., 2015).

6.3.2 Life cycle

To figure out how long the agents will maintain the BTAs, we first studied the life cycle distribution of the detected BTAs. Specifically, we set the time difference between the first and last tweets as the life cycle of a BTA. The first problem to address is determining whether a BTA is still active. In this study, we considered a BTA to be dead if it did not post any tweet in the past 180 days. Thus, we selected only the BTAs whose last tweet was posted 180 days before the end of our data collection time. Fig. 10a shows the distribution of the life cycle of the detected BTAs. We can see that over 50% of the BTAs were controlled by the agents over a year. This indicates that the agents would like to maximize their profits by maintaining their BTAs and re-use them for future customers. The reason might be that the registration cost is relatively high for a new BTA, because the platforms such as Weibo usually require users to bind a mobile phone number to their accounts. On the other hand, a well-maintained BTA is usually more human-like, which allows the agents to charge a higher price for the service provided by these higher-quality BTAs.

However, an interesting question arises: Why would the agents stop maintaining their BTAs? To answer this question, we computed the distribution of the abandoned time of the dead BTAs. We consider the timestamp of the last tweet in a dead BTA as the abandoned time. Fig. 10b shows the distribution of the abandoned time of the dead BTAs. The most interesting aspect of this distribution is that the agents began to abandon a large number of BTAs after September 2018. This phenomenon can be explained by the following three possible reasons: intensified supervision, competition between Weibo and other platforms, and the changes in supply and demand of this service within the Weibo platform. From the perspective of intensified supervision, China held a cybersecurity promotion week to intensify the crackdown on black industries in September 2018. The BTS as a form of black industry was also affected. From the perspective of the competition between Weibo and other platforms, the emergence of similar platforms, e.g., TikTok, affected the popularity of Weibo. For example, TikTok's global adoption was increased by over six times from 2017 to 2018, becoming the world's most downloaded App in the first half of 2018. Therefore, the agents were more likely to transfer part of their business from Weibo to other platforms. From the perspective of changes in the supply and demand of this

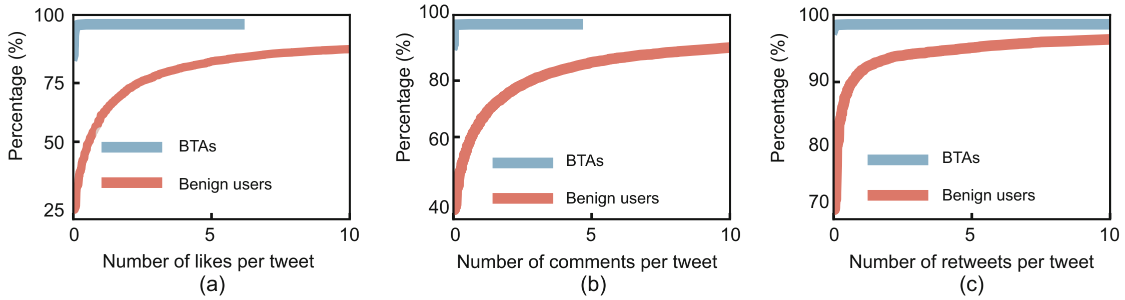

Fig. 9 Analysis of the influence of BTAs in social networks: (a) CDF of the number of likes per tweet; (b) CDF of the number of comments per tweet; (c) CDF of the number of retweets per tweet

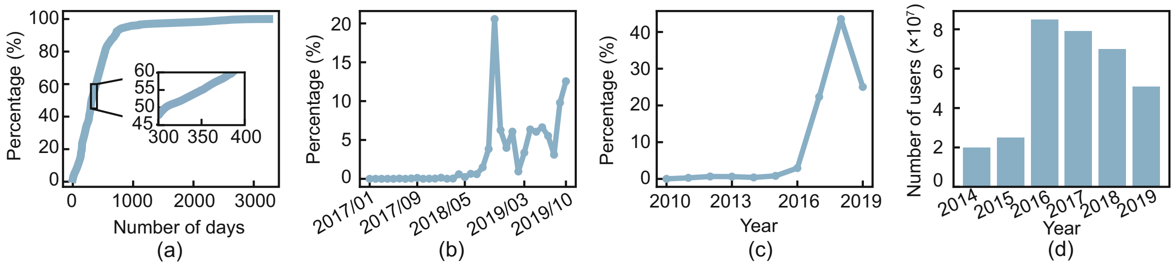

Fig. 10 BTA life cycle measurement: (a) CDF of BTA life cycles; (b) BTA abandoned time distribution; (c) BTA registration time distribution; (d) Weibo user increment distribution

service within the Weibo platform, we investigated the growth of new BTA registration and the usage growth of the Weibo platform. Specifically, we plotted the distribution of the number of newly registered BTAs per year. Note that we treat the timestamp of the first tweet of a BTA as its registration time. As shown in Fig. 10c, the registrations of new BTAs rapidly increased in 2016 and reached the peak in 2018. On the other hand, we plotted the distribution of the newly registered users per year on Weibo. As shown in Fig. 10d, we can see that the growth of the number of Weibo users has slowed down since 2016. Thus, the excess supply of BTAs and reduced demand for the BTS may cause further abandonment of BTAs.

6.3.3 Evasiveness

As discussed in Section 5.2, the BTAs are more advanced than the ones studied in previous works. To have a better understanding of their evasion tactics, we conducted a comprehensive analysis of their behaviors by examining their tweets. To facilitate our study, we divided the tweets of a BTA into bogus tweets and evasive tweets. The bogus tweets are the ones containing the target topics/hashtags, and the evasive tweets are the ones without the target topics/hashtags. These two kinds of tweets represent the attack and evasive behaviors, separately.

We first explored which behavior was the primary behavior of BTAs by plotting the CDF of the ratio of bogus tweets in a BTA. As shown in Fig. 11a, in about 50% of BTAs, the bogus tweet ratio was only 0.2. This means that the dominant behavior of BTAs is protecting themselves from being detected rather than performing attacks, because a smaller proportion of bogus tweets can help BTAs hide themselves better. It further indicates that the agents place greater importance on the protection of their BTAs. Second, we studied their switching patterns between the attack and evasive behaviors from two perspectives. The first perspective is quantitatively analyzing their attack behaviors by examining the number of continuous bogus tweets. We collected all the continuous bogus tweet phases from all the tweets posted by the detected BTAs. Then we counted the number of continuous bogus tweets in each phase and plotted their CDF in Fig. 11b. We can see that for 70% of the phases, the number of continuous bogus tweets was <10. Similarly, we analyzed the temporal property of their attack behaviors by examining the duration of

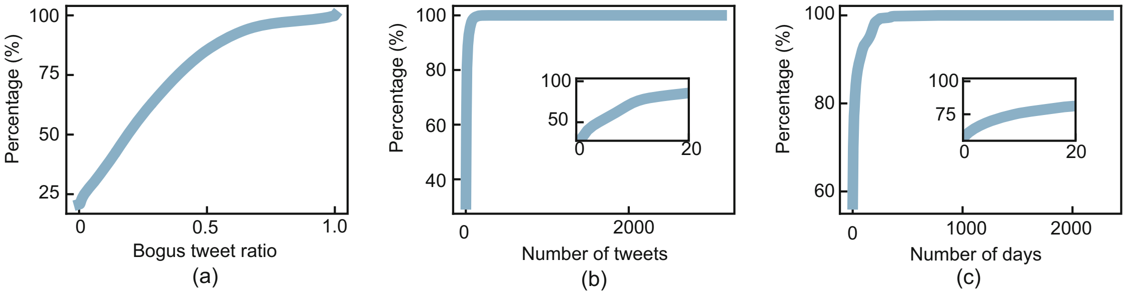

Fig. 11 Evasive behavior analysis: (a) ratio of bogus tweets; (b) number of continuous bogus tweets; (c) duration of continuous bogus tweets

each continuous bogus tweet phase. Fig. 11c shows the CDF of the duration. Nearly 75% of the continuous bogus tweet phases lasted <10 days. From the two perspectives above, we could see that the BTA attack behaviors are not continuous. Therefore, their attack behaviors are stealthy in each time period. According to what has been discussed above, we can infer that the agents may adopt strategies in which they divide the BTAs into several batches and carry out the attacks in turn.

To have a deeper understanding of the evasive tactics, we took a further look at the contents of their attack behaviors. We first analyzed the additional comments added by the BTAs in the bogus retweets. We found that the additional comments were very relevant to the original tweet. This motivated us to compute the semantic similarity between the additional comment and the original tweet in a retweet. Specifically, we used the TextRank (Mihalcea and Tarau, 2004) algorithm to extract the keywords in the original tweets and comments. Then we used the word2vec model (Mikolov et al., 2013; Le and Mikolov, 2014) pre-trained on the Chinese Wikipedia corpus to obtain the corresponding word embeddings of the extracted keywords. Finally, we calculated the cosine similarity of the embedding vectors between the comments and original tweets as the semantic similarity. We used the same method to calculate the semantic similarity in the retweets posted by benign users for comparison. As shown in Fig. 12, the similarity distribution of BTAs was almost identical to that of benign users. This observation supports the hypothesis that the agents may use some generative models to create the corresponding comments in the bogus retweets.

We also studied the content of BTAs' evasive tweets. We suspected that the agents might maintain a huge corpus, and each time BTAs post a tweet, they select a sentence from the corpus as the tweet. To test our hypothesis, we computed the number of appearances of each individual evasive tweet and plotted their CDF. As shown in Fig. 13, about 40% of the evasive tweets appeared more than once and nearly 25% of them appeared more than five times. This observation supports our hypothesis that the agents might maintain a huge corpus. To further determine the source of the corpus, we randomly sampled 1000 evasive tweets, and then manually searched each of them in the search engines to locate their source websites.

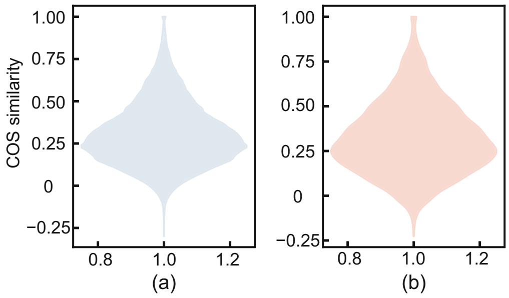

Fig. 12 Semantic similarity between the additional comments and the original tweets in the retweets: (a) BTAs; (b) benign users

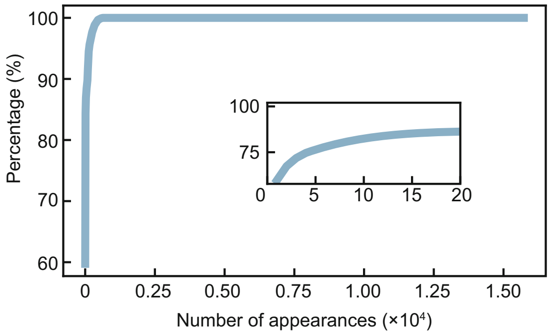

Fig. 13 CDF of the number of appearances of each individual evasive tweet

We divided the located source websites into eight types: quotations, forums, lyrics, poetries, news, games, landmarks, and others. From Table 4, we can roughly estimate that the quotations, news, and forum websites are the main sources. One possible reason could be that these websites are more popular and widely used than others.

Table 4 Source websites of evasive tweets

Source	Percentage (%)	Source	Percentage (%)
Quotations	23.6	Landmark	3.6
News	17.4	Lyrics	3.4
Forum	16.0	Game	3.2
Poetry	10.3	Others	22.5

7 Discussion

7.1 Limitations and future work

1. Potential bias. Our study is based on the agents we found and the BTAs we bought. They might not be the best representation of the whole population and could potentially bring bias to the study. Therefore, we are cautious about the conclusions drawn. For example, the target topics pushed by the detected BTAs might not be a good representation of all the involved topics. Thus, we do not make conclusions about the kind of topics that are more likely to involve BTSs. Instead, we list only the possible professions behind the discovered topics. To study the operating mechanism, the detected large-scale BTAs should be sufficient to support our conclusions. In addition, our conclusions about the mechanisms are based mostly on strong signals in the data. Therefore, the potential bias does not affect the conclusions made in our study.

2. Generalization. Our study focuses only on Weibo, the largest Chinese microblogging platform. However, our study is still valuable to society, because Weibo is now one of the most widely used micro-blogging platforms. At the end of 2019, Weibo had 516 million monthly active users and 222 million daily active users, while Twitter had 152 million monetizable daily active users in December 2019.

The methodological framework proposed within this research holds potential for application across diverse platforms, contingent upon the existence of linguistic feature differences among user accounts on these platforms. To substantiate the generalization of our approach, we extended the evaluation to the Cresci-2017 dataset (Cresci et al., 2017), a well-known Twitter bot dataset, encompassing 3474 benign users and 3351 fake followers. We removed the user profiles without a description, leaving 3095 benign users and 2278 fake followers. At a high level, our approach deployed linguistic features extracted from user descriptions and amalgamated these with five profile-based features.

To extract the linguistic features, we randomly selected 1000 descriptions from benign users and fake followers, forming a training set for a BERT model. The BERT model's objective was to distinguish between descriptions of benign users and those of fake followers. By applying the BERT model to the remaining descriptions, we generated a fake follower detection dataset, merging the prediction results with five profile-based features: statuses_count, followers_count, friends_count, favourites_count, and listed_count. We divided the dataset based on a 7:3 ratio for training and test and trained an XGBoost classifier on the training set. The final classifier demonstrated a precision of 99.5% and a recall of 99.4%, signifying the successful extrapolation of our methodology in other OSNs.

In future studies, we envision extending the evaluation of the BTS ecosystem to other microblogging platforms, such as Twitter.

7.2 Suggestions

1. Platform. The results in our study indicate that the evasive tactics adopted by the agents are able to bypass the detection of Weibo's anti-spam system. The primary reason is that the BTAs are disguised well by the agents. Their behaviors look benign superficially because they do not influence their neighbors in the network but simply tweet and retweet as humans do. Furthermore, the content of their tweets is highly human-like. We have reported our findings to Weibo officials and are waiting for their reply. We hope that Weibo officials will pay more attention to this new kind of black service and refer to our findings to fortify their anti-spam systems.

2. Society. In our research, we reveal that BTS is a more advanced and evasive illicit service on Weibo. Once it succeeds, a large number of users will be affected. Even worse, users might be unaware that they are being misled by the fake topics/news

promoted by the BTS. Therefore, we suggest that Weibo users need to be vigilant about the HTP contents.

7.3 Ethical issues

To conduct our study, we purchased BTSs from the agents and crawled data from Weibo. Our purchase behavior might be misunderstood as promoting the growth of the underground economy. To mitigate this negative impact, we carefully conducted our experiments to minimize the revenue and impact on Weibo while maximizing the research results. In the process, we made sure that any interactions with the underground market would not harm the normal users of Weibo. For example, we issued a statement on our honeypot account to tell users that the purpose of our account was to study BTS and not to trust the fans of our account, which can prevent users from being misled by this account. In the user information crawling process, we strictly followed Weibo's API instructions and did not make excess calls to Weibo. In addition, the information we collected is publicly available and does not intrude on users' privacy.

7.4 Adaptive agents

Despite the potential for adaptive agents to understand our detection methodology, their circumvention of the algorithm remains a significant challenge, attributable to two key factors. First, the detection approach leverages the evasive tactics of BTAs as a basis for their identification. If the BTAs want to evade our detection method, they must forego their evasion tactics. This action will inadvertently render the BTAs to be more susceptible to recognition as conventional bots. Therefore, the accounts controlled by them will be easier to recognize by humans or platforms. Further, the absence of their evasive tactics reduces the camouflaging potential of the bogus traffic attack, rendering it to be transparent as a bot activity and thereby negating its efficacy. Second, in compliance with the principle of security through obscurity, we withhold open-source access to the parameters of our customized BERT model. This strategic decision impedes adaptive agents from conducting an effective adversarial attack against our model.

8 Conclusions

BTSs are mature and advanced malicious services that have the potential to mislead public opinion. To the best of our knowledge, this work reports the first systematic study of this service, including the analysis of BTS marketplaces, detection of BTS attacks, and disclosure of the BTS operating mechanism.

To answer the question about the BTS marketplace, we investigated the BTS marketplace and contributed a ground-truth dataset of BTAs to facilitate future research on this topic. We found that the BTS agents can be found through various channels, including search engines, e-commerce platforms, and the OSN itself. We also found that the BTS agents are more likely to use IM Apps as the communication channel.

To answer the question about the detection of bogus traffic, we used the ground-truth dataset to identify a strong signal to detect BTAs as a bridge to identify bogus traffic, which is the linguistic difference between evasive tweets and normal tweets. The results showed that the customized BERT model with the XGBoost classifier can effectively detect BTAs.

To answer the question about the BTS operating mechanism, we revealed the operating mechanism from the attack cycle side and the attack entity side. The BTS attack cycle is typically <10 days. Within every attack cycle, most of the bogus traffic is concentrated at the beginning and the end. On the attack entity side, we found that BTAs are harmless to social connections and have a relatively long life cycle. The reasons behind the abundance of BTAs may come from the intensified supervision and the changes in the supply and demand for this service. Finally, we found that BTAs adopt various evasive tactics, including massively tweeting evasive tweets and retweeting bogus retweets with semantically similar comments.

Contributors

Ping HE designed the research, processed the data, conducted the experiments, and drafted the paper. Xuhong ZHANG, Changting LIN, Ting WANG, and Shouling JI helped organize the paper. Ping HE and Shouling JI revised and finalized the paper.

Acknowledgements

The authors would like to thank Haofei YU for suggestions on the detection method and Xueyan LYU for the investigation of the marketplace of BTS. The authors would also like to thank Tianyu DU and Yiming WU for their suggestions to revise the paper. We thank the support from the SRTP project in the College of Computer Science and Technology of Zhejiang University, and the NGICS platform of Zhejiang University.

Conflict of interest

Shouling JI is a corresponding expert of *Frontiers of Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Ali Alhosseini S, Bin Tareaf R, Najafi P, et al., 2019. Detect me if you can: spam bot detection using inductive representation learning. Companion Proc World Wide Web Conf, p.148-153. https://doi.org/10.1145/3308560.3316504

Alibaba Inc., 2020. Alibaba Annual Report. https://static.alibabagroup.com/reports/fy2020/ar/ebook/en/index.html [Accessed on Feb. 23, 2022].

Alvisi L, Clement A, Epasto A, et al., 2013. SoK: the evolution of sybil defense via social networks. IEEE Symp on Security and Privacy, p.382-396. https://doi.org/10.1109/SP.2013.33

Beskow DM, Carley KM, 2019. Its all in a name: detecting and labeling bots by their name. *Comput Math Organ Theory*, 25(1):24-35. https://doi.org/10.1007/s10588-018-09290-1

Beskow DM, Carley KM, 2020. You are known by your friends: leveraging network metrics for bot detection in Twitter. In: Tayebi MA, Glässer U, Skillicorn DB (Eds.), Open Source Intelligence and Cyber Crime: Social Media Analytics. Springer, Switzerland, p.53-88. https://doi.org/10.1007/978-3-030-41251-7_3

Booij TM, Verburgh T, Falconieri F, et al., 2021. Get rich or keep tryin' trajectories in dark net market vendor careers. IEEE European Symp on Security and Privacy Workshops, p.202-212. https://doi.org/10.1109/EuroSPW54576.2021.00028

Boshmaf Y, Logothetis D, Siganos G, et al., 2015. Integro: leveraging victim prediction for robust fake account detection in OSNs. Network and Distributed System Security Symp, p.8-11. https://doi.org/10.14722/ndss.2015.23260

Cao Q, Yang XW, Yu JQ, et al., 2014. Uncovering large groups of active malicious accounts in online social networks. Proc ACM SIGSAC Conf on Computer and Communications Security, p.477-488. https://doi.org/10.1145/2660267.2660269

Chen TQ, Guestrin C, 2016. XGBoost: a scalable tree boosting system. https://doi.org/10.48550/arXiv.1603.02754

Cresci S, di Pietro R, Petrocchi M, et al., 2017. The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. Proc 26th Int Conf on World Wide Web Companion, p.963-972. https://doi.org/10.1145/3041021.3055135

Cresci S, Petrocchi M, Spognardi A, et al., 2019. On the capability of evolved spambots to evade detection via genetic engineering. *Online Soc Netw Med*, 9:1-16. https://doi.org/10.1016/j.osnem.2018.10.005

Cuevas A, Miedema F, Soska K, et al., 2022. Measurement by proxy: on the accuracy of online marketplace measurements. 31st USENIX Security Symp, p.2153-2170.

de Cristofaro E, Friedman A, Jourjon G, et al., 2014. Paying for likes? Understanding Facebook like fraud using honeypots. Proc Conf on Internet Measurement Conf, p.129-136. https://doi.org/10.1145/2663716.2663729

Devlin J, Chang MW, Lee K, et al., 2018. BERT: pre-training of deep bidirectional Transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805

Dutta HS, Chakraborty T, 2020. Blackmarket-driven collusion among retweeters—analysis, detection, and characterization. *IEEE Trans Inform Forens Secur*, 15:1935-1944. https://doi.org/10.1109/TIFS.2019.2953331

Elmas T, Overdorf R, Özkalay AF, et al., 2021. Ephemeral astroturfing attacks: the case of fake Twitter trends. IEEE European Symp on Security and Privacy, p.403-422. https://doi.org/10.1109/EuroSP51992.2021.00035

Feng SB, Wan HR, Wang NN, et al., 2021. TwiBot-20: a comprehensive Twitter bot detection benchmark. Proc 30th ACM Int Conf on Information & Knowledge Management, p.4485-4494. https://doi.org/10.1145/3459637.3482019

Feng SB, Tan ZX, Li R, et al., 2022. Heterogeneity-aware Twitter bot detection with relational graph transformers. *Proc AAAI Conf Artif Intell*, 36(4):3977-3985. https://doi.org/10.1609/aaai.v36i4.20314

Feng SB, Tan ZX, Wan HR, et al., 2023. TwiBot-22: towards graph-based Twitter bot detection. https://doi.org/10.48550/arXiv.2206.04564

Freitas C, Benevenuto F, Ghosh S, et al., 2015. Reverse engineering socialbot infiltration strategies in Twitter. IEEE/ACM Int Conf on Advances in Social Networks Analysis and Mining, p.25-32. https://doi.org/10.1145/2808797.2809292

Guo ZY, Wang LQ, Wang YF, et al., 2018. Public opinion spamming: a model for content and users on Sina Weibo. Proc 10th ACM Conf on Web Science, p.210-214. https://doi.org/10.1145/3201064.3201104

HuggingFace, 2022. BERT Base Chinese Model. https://huggingface.co/bert-base-chinese [Accessed on May 26, 2022].

Jakesch M, Garimella K, Eckles D, et al., 2021. Trend alert: a cross-platform organization manipulated Twitter trends in the Indian general election. *Proc ACM Human-Computer Interact*, 5(CSCW2):379. https://doi.org/10.1145/3479523

JD Inc., 2020. JD Annual Report. https://ir.jd.com/static-files/fc93d5dd-9437-4141-9191-f960ba46874b [Accessed on May 26, 2022].

Just MR, Crigler AN, Metaxas P, et al., 2012. "It's trending on Twitter"—an analysis of the Twitter manipulations in the Massachusetts 2010 Special Senate Election. Annual Meeting of the American Political Science Association.

Le QV, Mikolov T, 2014. Distributed representations of sentences and documents. https://arxiv.org/abs/1405.4053

Liu PF, Yuan WZ, Fu JL, et al., 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*, 55(9):195. https://doi.org/10.1145/3560815

Mihalcea R, Tarau P, 2004. TextRank: bringing order into text. Proc Conf on Empirical Methods in Natural Language Processing, p.404-411. https://aclanthology.org/W04-3252

Mikolov T, Chen K, Corrado G, et al., 2013. Efficient estimation of word representations in vector space. https://arxiv.org/abs/1301.3781

PDD Inc., 2020. PDD Annual Report. https://investor.pddholdings.com/static-files/0ad89f79-7123-4072-8662-d5509227526c [Accessed on May 26, 2022].

Song J, Lee S, Kim J, 2015. CrowdTarget: target-based detection of crowdturfing in online social networks. Proc 22nd ACM SIGSAC Conf on Computer and Communications Security, p.793-804. https://doi.org/10.1145/2810103.2813661

Stringhini G, Wang G, Egele M, et al., 2013. Follow the green: growth and dynamics in Twitter follower markets. Proc Conf on Internet Measurement Conf, p.163-176. https://doi.org/10.1145/2504730.2504731

Thomas K, McCoy D, Grier C, et al., 2013. Trafficking fraudulent accounts: the role of the underground market in Twitter spam and abuse. Proc 22nd USENIX Conf on Security, p.195-210. https://dl.acm.org/doi/10.5555/2534766.2534784

Thomas K, Li F, Grier C, et al., 2014. Consequences of connectivity: characterizing account hijacking on Twitter. Proc ACM SIGSAC Conf on Computer and Communications Security, p.489-500. https://doi.org/10.1145/2660267.2660282

Torres-Lugo C, Yang KC, Menczer F, 2022. The manufacture of partisan echo chambers by follow train abuse on Twitter. *Proc Int AAAI Conf Web Soc Med*, 16(1):1017-1028. https://doi.org/10.1609/icwsm.v16i1.19354

van Wegberg R, Tajalizadehkhoob S, Soska K, et al., 2018. Plug and prey? Measuring the commoditization of cybercrime via online anonymous markets. Proc 27th USENIX Conf on Security Symp, p.1009-1026. https://doi.org/10.5555/3277203.3277279

Weerasinghe J, Flanigan B, Stein A, et al., 2020. The pod people: understanding manipulation of social media popularity via reciprocity abuse. Proc Web Conf, p.1874-1884. https://doi.org/10.1145/3366423.3380256

Woolley SC, 2016. Automating power: social bot interference in global politics. *First Mond*, 21(4). https://doi.org/10.5210/fm.v21i4.6161

Yang C, Harkreader R, Gu GF, 2013. Empirical evaluation and new design for fighting evolving Twitter spammers. *IEEE Trans Inform Forens Secur*, 8(8):1280-1293. https://doi.org/10.1109/TIFS.2013.2267732

Yu HF, Kaminsky M, Gibbons PB, et al., 2006. SybilGuard: defending against sybil attacks via social networks. *SIGCOMM Comput Commun Rev*, 36(4):267-278. https://doi.org/10.1145/1151659.1159945

Yu HF, Gibbons PB, Kaminsky M, et al., 2010. SybilLimit: a near-optimal social network defense against sybil attacks. *IEEE/ACM Trans Netw*, 18(3):885-898. https://doi.org/10.1109/TNET.2009.2034047

Yuan D, Miao YL, Gong NZ, et al., 2019. Detecting fake accounts in online social networks at the time of registrations. Proc ACM SIGSAC Conf on Computer and Communications Security, p.1423-1438. https://doi.org/10.1145/3319535.3363198

Zhang YB, Ruan X, Wang HN, et al., 2017. Twitter trends manipulation: a first look inside the security of Twitter trending. *IEEE Trans Inform Forens Secur*, 12(1):144-156. https://doi.org/10.1109/TIFS.2016.2604226

Zheng HZ, Xue MH, Lu H, et al., 2017. Smoke screener or straight shooter: detecting elite sybil attacks in user-review social networks. https://arxiv.org/abs/1709.06916

List of supplementary materials

1 Communication channel analysis
2 Honeypot account
3 Evasive tweets
4 Profile-based features
5 Case study
6 Weibo authentication rules
Fig. S1 The announcement in our honeypot account
Fig. S2 Two examples of evasive tweets in our dataset
Fig. S3 Feature analysis of the profile-based features
Fig. S4 Bogus traffic distributions for three superstars