Detecting Inactive Cyberwarriors from Online Forums

Ruei-Yuan Wang
Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
zx57680@gmail.com

Hung-Hsuan Chen
Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
hhchen1105@acm.org

***Abstract*—The proliferation of misinformation has emerged as a new form of warfare in the information age. This type of warfare involves cyberwarriors, who deliberately propagate messages aimed at defaming opponents or fostering unity among allies. In this study, we investigate the level of activity exhibited by cyberwarriors within a large online forum, and remarkably, we discover that only a minute fraction of cyberwarriors are active users. Surprisingly, despite their expected role of actively disseminating misinformation, cyberwarriors remain predominantly silent during peacetime and only spring into action when necessary. Moreover, we analyze the challenges associated with identifying cyberwarriors and provide evidence that detecting inactive cyberwarriors is considerably more challenging than identifying their active counterparts. Finally, we discuss potential methodologies to more effectively identify cyberwarriors during their inactive phases, offering insights into better capturing their presence and actions. The experimental code is released for reproducibility: https://github.com/Ryaninthegame/Detect-Inactive-Spammers-on-PTT.**

***Index Terms*—cyber attack, graphical neural network, forum, spammer, netizen, information warfare, media framing, filter bubble, cyberwarrior**

I. Introduction

Social media has emerged as a crucial platform for information sharing, leading politicians, political parties, and governments to enlist the services of public relations (PR) companies and social media curators to bolster their online reputations. Regrettably, these PR firms occasionally engage in the deliberate dissemination of plausible but potentially incorrect or partially accurate statements on the Internet, employing techniques such as spin control or media framing. A prominent example of this phenomenon is Russia's interference in the 2016 United States election, with propaganda estimated to have reached 126 million Facebook users and over 20 million Instagram users [1].

Online propaganda typically relies on a multitude of user accounts to spread information and create a false impression of the formation of public opinion. These accounts, referred to as "cyberwarriors" in this paper, can be generated automatically or purchased at an affordable cost online. For example, a Chinese government document in 2021 reveals that accessing hundreds of active Facebook and Twitter accounts costs 5000 RMB per month, approximately 710 USD [2].

This work is partially supported by the National Science and Technology Council of Taiwan under grant 110-2222-E-008-005-MY3.

TABLE I
A comparison of the AUPRC scores of detecting active and inactive cyberwarriors using different machine learning models. The results show that detecting inactive cyberwarriors is much more challenging.

	active users	inactive users	diff
XGBoost	0.8892	0.5157	0.3735
LightGBM	0.7421	0.4888	0.2533
Random Forest	0.8317	0.5147	0.3163

The detection of cyberwarriors plays a pivotal role in combating the propagation of fake information. Previous studies have explored the behaviors of suspicious accounts and spammers on various platforms, proposing methodologies to detect them [3], [4], [5], [6], [7], [8]. However, most of these studies focus on highly active users who exhibit extensive engagement on the platform, such as leaving comments, sharing photos, and initiating discussions. Apparently, detecting active spammers based on abundant activity logs is comparatively straightforward.

Perhaps to the surprise of many people, although the mission of a cyberwarrior is to disseminate messages, a cyberwarrior account may remain inactive for an extended period before disseminating misleading posts [9]. Consequently, active cyberwarriors make up only a tiny proportion of the total cyberwarriors. As a result, identifying inactive cyberwarriors may pose a significantly greater challenge. To validate this point, we conducted a preliminary study demonstrating the ease of detecting spammers among active users using supervised learning techniques and the difficulty of detecting inactive spammers. As illustrated in Table I, when applying some of the most successful machine learning models (XGBoost, LightGBM, and Random Forest) to detect spammers among active users, we achieve decent scores. However, these numbers decrease significantly when targeting inactive users, with an average drop of over 30% in the area under the precision-recall curve (AUPRC).

In this paper, our aim is to quantify a user's level of activeness and focus on identifying abnormal accounts among inactive users. Since inactive users provide limited activity logs

as features, we enhance the available clues by incorporating social information from two perspectives. First, we use data from inactive users' connected accounts to generate social-related features. Second, we employ graph neural networks (GNNs) as training models to capture the relationships between different accounts. Consequently, even if an account exhibits limited activity logs, we can leverage information from its neighboring accounts to detect its status (normal or spammer). Our findings indicate that these simple strategies significantly improve the effectiveness of discovering inactive spammers while also slightly enhancing the detection of active spammers.

In summary, this paper makes the following contributions.

- We demonstrate that detecting spammers among inactive users is considerably more challenging than among active users. Additionally, we highlight that a substantial number of spammers are inactive users, which has not received significant attention in previous studies that primarily focus on active users.
- We introduce social-related features and employ graph neural network models to leverage information from an account's neighboring accounts. Through comprehensive experiments, we demonstrate that these simple strategies improve the detection of inactive and active spammers compared to other baseline methods.
- For reproducibility purposes, we release the experimental dataset and the accompanying code. In addition, the dataset can serve as a valuable benchmark for spammer detection, as administrators of a large forum have manually labeled the spammers in our dataset.

The rest of the paper is organized as follows. Section II reviews previous studies on spammer detection. In Section III, we present the statistics and features of our studied forum. In Section IV, we analyze the behaviors of active and inactive spammers and compare the challenges associated with their detection. Finally, we conclude and discuss our work in Section V.

II. Related Work

Detecting abnormal accounts has been a topic of extensive research, with various approaches and techniques employed to establish the relationship between account features and their classification as normal or abnormal. Machine learning models have proven to be valuable tools in this regard [3], [4], [5], [6], [7], [8]. These models leverage relevant information, such as an account's public profile and behaviors, to identify patterns indicative of abnormal activity. Although these machine learning methods have demonstrated promising performance in determining the status of active accounts, they often encounter difficulties when dealing with less active accounts, as evidenced by our preliminary study in Table I. The limited information left by inactive accounts poses challenges for feature extraction and classification, leading to suboptimal performance in detecting such accounts.

Graphs are a natural choice for modeling relationships due to their ability to represent complex interactions and connections between entities in a visually intuitive and versatile manner [10], [11], [12], [13]. In recent years, graph neural networks have received significant attention due to their remarkable performance in various domains, including biomedical research, social network analysis, and abnormal sensor detection [14], [15], [16]. Exploiting the inherent graph structure of social networks to detect abnormal accounts becomes a natural choice. Consequently, researchers have leveraged graph-based approaches [17] and, more recently, graph neural networks (GNNs) [18], [19], [20] to model social networks and make predictions. These techniques can effectively identify suspicious patterns and uncover abnormal behavior by capturing relational information among accounts. However, despite the progress made in this field, detecting abnormal accounts remains challenging, particularly when dealing with less active or inactive accounts. The complexities that arise from the limited availability of information and the evolving nature of suspicious behavior necessitate further research and the development of advanced techniques to enhance detection accuracy and robustness.

III. The PTT Forum

We collect the experimental dataset from the PTT forum. In this section, we provide a comprehensive introduction to the PTT forum, including its background, statistics, and noteworthy features, to familiarize readers with the platform.

A. Introduction of PTT

PTT, established in 1995, has emerged as one of the largest and most influential forums in Taiwan. With a massive user base that exceeds 1.5 million registered accounts and encompasses over 20,000 discussion boards covering a wide range of topics, PTT serves as a vibrant platform where users actively engage in discussions, sharing insights, opinions, and experiences [21]. The popularity of the forum is evidenced by the staggering volume of user-generated content, with an average daily production of more than 20,000 articles and over 500,000 comments.

The PTT forum caters to the diverse interests of Taiwanese citizens, providing an avenue for discussions on a myriad of subjects, including shopping experiences, celebrity gossip, news updates, religions, movies, life goals, and notably critical societal events. In particular, the platform has played an important role in facilitating discussions during key historical moments in Taiwan in the last decades. For instance, during the Sunflower Student Movement in 2014, which involved a three-week occupation of Taiwan's Legislative Yuan[1] by civic groups and students, a single discussion board on PTT witnessed the simultaneous presence of over 100,000 users, which further encouraged more citizens to join the movement. This event demonstrates the forum's ability to mobilize individuals and foster engagement. Similarly, during the 2016 presidential election and the 2018 city mayor election, PTT attracted similar numbers of users concurrently visiting a discussion

[1]Logislative Yuan is the unicameral legislature of Taiwan, similar to UK Parliament and US Congress.

board, further highlighting its relevance and impact in shaping public discourse.

Given the substantial influence of PTT, various entities, such as journalists, politicians, political parties, and the entertainment industry, actively monitor the platform to gauge public opinion. In particular, politicians recognize the importance of securing votes, particularly from the young and middle-aged demographics, by leveraging PTT as a battleground to connect with potential supporters and address their concerns.

PTT stands out among other online forums due to its distinctive features and mechanisms. One notable feature is its commenting system, where users can express their sentiment towards an article through options such as liking, disliking, or remaining neutral, accompanied by a 45-character comment. Moreover, articles that receive a significant number of likes or dislikes are visually highlighted with special colored symbols, capturing users' attention and potentially triggering further engagement. This feedback loop reinforces the amplification of likes or dislikes and subsequently increases the visibility of such articles, leading to increased exposure and potential impact.

However, with the substantial influence and visibility of PTT, there have been instances where politicians, political parties, and public relations (PR) firms resort to disseminating disinformation on the platform for various purposes, including media framing, attacking opponents, or self-promotion. The unique highlighting system introduced earlier serves as a motivation for individuals with specific agendas to mobilize accounts and accumulate a large number of likes or dislikes on selected articles in a short period of time, aiming to generate further attention around these topics [22]. These dynamics present challenges in distinguishing between genuine user participation and orchestrated manipulations, underscoring the need for robust detection mechanisms.

B. Experimental Data Collection on PTT

To conduct our research, we collected experimental data from the PTT forum, focusing on a specific time period and a subset of accounts associated with suspicious activities. From March 2019 to December 2019, PTT officially announced $7,581$ accounts as spammers, primarily suspected of attempting to influence the city mayor elections of six major cities in Taiwan in November 2018 and the upcoming presidential election in March 2020. Out of these $7,581$ accounts, $4,918$ of them have at least one activity record related to article posting or commenting. However, it is worth noting that most of these $4,918$ accounts exhibit minimal activity, with up to 92% of them having no more than 0.18 activities per day, indicating a high degree of inactivity. Consequently, relying solely on the activity logs of these accounts to detect whether they are spammers or regular users, as suggested by previous studies, may not yield optimal results.

To capture the relevant data for our analysis, we crawled the articles from July 1, 2018, to December 29, 2019, based on the following considerations. First, the PTT announced the first batch of suspicious accounts in March 2019, approximately four months after the city mayors' election on November 24, 2018. Given this timeline, we assume that these accounts began their actions approximately six months prior to the election. Therefore, we started our data collection on July 1, 2018, to include the crucial period leading up to the election. Secondly, the PTT announced its last batch of suspicious accounts and suspended their posting and commenting permissions on December 29, 2019. Consequently, we set this date as the final crawling day to ensure comprehensive coverage of relevant data.

After crawling the articles and comments, we discovered that the total number of associated accounts (that is, including the authors and commentors) exceeded $200,000$, which would require substantial memory space, particularly when employing graph neural networks (details of which will be introduced in Section IV). To manage the dataset more effectively, we further pruned the articles based on specific criteria. First, we included only articles with at least 90 comments, ensuring a reasonable level of engagement for comprehensive analysis. Second, if the associated accounts of an article contained fewer than three spammers, we excluded the article, focusing on those articles where suspicious activities were more prevalent. Finally, we include a maximum of 80 commentors for the remaining articles. Specifically, if the number of spammers associated with an article was less than 80, we included all the spammers; we included regular users in chronological order until we reached a total of 80 accounts. On the contrary, if more than 80 spammers were associated with an article, we selected the earliest 80 spammers while excluding regular user accounts. Following these criteria, we collected a dataset consisting of $44,602$ user accounts, with 912 of them identified as spammers by PTT administrators.

All subsequent experiments and analyses presented in this study are based on the pruned dataset obtained after the selection process.

IV. Analysis

This section presents the empirical activeness scores of cyberwarriors and compares the effectiveness of different models to detect them. It provides insights into the activity levels of spammers and explores the performance of various algorithms in identifying them.

A. Most spammers are less active than normal users

We define the degree of activeness of an account by considering the average number of daily articles and comments. To assess the activity levels of the collected users, we calculate the active value for each user and categorize them into 10 groups, denoted G_1 to G_{10}. Each group G_i contains users whose active values fall within the $(i-1)$th percentile and the ith percentile among all users.

Table II presents the number of normal and spammer accounts in each group G_i. As evident from the column "# normal accounts" and the column "CDF of normal accounts", the number of normal accounts remains relatively consistent across the groups. However, the activeness values of spammers

TABLE II
THE NUMBER OF NORMAL USERS AND SPAMMERS FOR EACH GROUP. THE SYMBOL $[p, q)$ REFERS TO THE PERCENTILE OF ACTIVE VALUE r IN THE RANGE: $p \leq r < q$.

Group	Percentile of active value	Active value	# normal accounts	CDF of normal accounts (a)	# spammers	CDF of spammers (b)	(b) − (a)
G_1	[0%, 10%)	0-18	4112	9%	222	24%	15%
G_2	[10%, 20%)	19-45	4418	20%	163	42%	22%
G_3	[20%, 30%)	46-84	4508	30%	86	52%	22%
G_4	[30%, 40%)	85-135	4223	40%	59	58%	18%
G_5	[40%, 50%)	136-211	4453	50%	57	64%	14%
G_6	[50%, 60%)	212-315	4096	59%	76	73%	14%
G_7	[60%, 70%)	316-494	4320	69%	112	85%	16%
G_8	[70%, 80%)	495-817	4368	79%	67	92%	13%
G_9	[80%, 90%)	818-1663	4638	90%	51	98%	8%
G_{10}	[90%, 100%]	≥ 1664	4554	100%	19	100%	0%

TABLE III
THE AUPRC SCORES OF VARIOUS NON-GNN MODELS (WITHOUT SOCIAL FEATURES). WE REPEAT EACH EXPERIMENT 10 TIMES AND REPORT THE MEAN ± STANDARD DEVIATION.

	[0%, 10%)	[10%, 20%)	[80%, 100%]
XGBoost	0.52 ± 0.01	0.48 ± 0.03	0.89 ± 0.01
LightGBM	0.49 ± 0.02	0.40 ± 0.04	0.74 ± 0.02
Random Forest	0.51 ± 0.03	0.27 ± 0.02	0.83 ± 0.02
Fully Connected	0.35 ± 0.06	0.38 ± 0.05	0.75 ± 0.03
ConvNet	0.17 ± 0.06	0.26 ± 0.14	0.80 ± 0.33
Soft Voting [22]	0.40 ± 0.01	0.43 ± 0.01	0.76 ± 0.01
Hard Voting [22]	0.43 ± 0.02	0.47 ± 0.02	0.70 ± 0.03
Stacking [22]	0.42 ± 0.01	0.47 ± 0.03	0.67 ± 0.01

exhibit a significant skew. As indicated in the last column of Table II, the cumulative distribution function (CDF) of spammers for each row consistently exceeds the CDF of regular accounts. This implies that, compared to normal users, most spammers exhibit lower activity levels.

Since cyberwarriors are expected to disseminate information, it may be argued that cyberwarriors should demonstrate higher levels of activity. Thus, our empirical observation – spammers are typically less active during non-conflict periods – may be a surprise to many people. However, we found that previous research on Twitter accounts aligns with our findings and supports the claim that cyberwarriors often exhibit extended periods of inactivity during peacetime and only engage in extensive posting when necessary [9].

B. Supervised learning is successful in detecting active spammers, but not inactive spammers

Given that spammers are generally less active than normal users, detecting them may pose a greater challenge for algorithms because of the limited clues they leave behind.

To validate this conjecture, we selected various algorithms and tested their effectiveness in identifying active and inactive spammers. The algorithms include two popular algorithms known for their success in Kaggle competitions (XGBoost and LightGBM), deep learning models such as fully connected networks and convolutional neural networks (ConvNet), and recently proposed approaches for spammer detection for PTT, namely soft voting, hard voting, and stacking ensemble [22]. For each account a, we considered three features. First, we computed the average popularity of a user's associated articles (i.e., the account a's posted or commented articles). In particular, we computed the total number of comments for all m articles and divided by m. Second, we calculated the average sentiment of comments about articles by subtracting the number of dislikes from the number of likes for each of the m articles and computing the average. These two features were included because previous studies indicate that spammers often generate many comments on selected articles to increase their visibility. Lastly, we incorporated the active period of an account as the third feature.

To evaluate the performance, we used the area under the precision-recall curve (AUPRC) as the metric. Given the highly imbalanced nature of our data set, with a percentage of spammers ranging from 0.4% to 5% in each group (as shown in Table II), AUPRC was considered more appropriate than the area under the receiver operating characteristic curve (AUROC). AUROC tends to overstate the performance of a classifier when the positive class is the minority class, potentially leading to misleading results [23], [24]. In contrast, AUPRC is suitable for scenarios where the positive class is of interest and represents the minority, as it accounts for precision and recall without considering true negatives (i.e., the negative instances that are predicted as negative by a model).

Table III reports the AUPRC scores of the selected algorithms for three groups based on users' activeness values: $[0, 10\%)$, $[10\%, 20\%)$, and $[80\%, 100\%]$. The results reveal that as the activeness value increases (i.e., in the $[80\%, 100\%]$ group), the improved scores of the average AUPRC range from 20% to 63% compared to the users in the $[0\%, 10\%)$ or $[10\%, 20\%)$ groups. This finding aligns with our hypothesis that supervised learning algorithms excel at detecting active cyberwarriors. However, identifying inactive cyberwarriors is significantly more challenging. Unfortunately, most cyberwarriors exhibit low activity during peacetime, making it possible to identify them only when they engage in aggressive posting and sharing of articles.

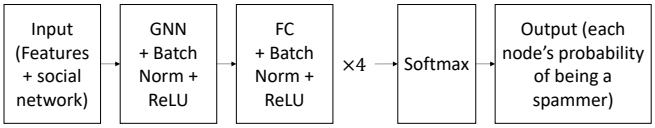

Fig. 1. The structure of the GNN-based models. The GNN is either GCN, TAGCN, or GAT.

TABLE IV
THE AUPRC SCORES OF VARIOUS GNN MODELS AND THE BEST NON-GNN MODEL (WITHOUT SOCIAL FEATURES). WE REPEAT EACH EXPERIMENT 10 TIMES AND REPORT THE MEAN ± STANDARD DEVIATION.

	[0%, 10%)	[10%, 20%)	[80%, 100%]
XGBoost	0.52 ± 0.01	0.48 ± 0.03	0.89 ± 0.01 †
GCN	$\mathbf{0.66} \pm 0.18$	0.38 ± 0.13	0.72 ± 0.07
TAGCN ($K = 1$)	$\mathbf{0.64} \pm 0.04$	$\mathbf{0.79} \pm 0.06$	$\mathbf{0.89} \pm 0.07$ †
TAGCN ($K = 2$)	$\mathbf{0.68} \pm 0.02$	$\mathbf{0.84} \pm 0.05$ †	$\mathbf{0.89} \pm 0.08$ †
TAGCN ($K = 3$)	$\mathbf{0.71} \pm 0.04$ †	$\mathbf{0.80} \pm 0.07$	$\mathbf{0.89} \pm 0.06$ †
GAT	$\mathbf{0.62} \pm 0.09$	$\mathbf{0.77} \pm 0.05$	$\mathbf{0.89} \pm 0.06$ †

C. Social information helps discover inactive spammers

This section demonstrates that integrating social information can enhance the identification of inactive spammers. We explore two perspectives for incorporating social information: utilizing graphical neural network (GNN) models and designing specialized social features. Our experimental results validate the effectiveness of both approaches.

1) GNN models: This section introduces GNN-based models and explains how we construct a social network. Leveraging the social information embedded in the network, models can potentially extract valuable insights from neighboring accounts, even when an inactive account provides limited activity logs.

We consider each account as a node in a graph, connecting two nodes with an edge if the corresponding accounts co-appear in an article (either as commentors or with one as the poster and the other as a commentor). To represent the graph, we generate an adjacency matrix $A = [a_{i,j}]_{i,j=1,\ldots,n}$, where n denotes the number of nodes, and $a_{i,j} = 1$ if there exists an edge connecting nodes i and j, and 0 otherwise.

We selected three representative GNN models as part of our learning framework: graph convolutional networks (GCN) [25], topology adaptive graph convolutional networks (TAGCN) [26], and graph attention network (GAT) [27]. These GNNs incorporate information from neighboring nodes into each node i through recursive information propagation, thereby fusing the neighboring information with node i's information. The differences among these models lie in the range of the neighboring area and the mechanism used to integrate information. Figure 1 provides an overview of the structure of the neural network with GNN models.

Table IV compares the best non-GNN model (XGBoost) with GNN-based models. Most models perform satisfactorily in detecting spammers from active accounts, as indicated in the last column. However, when targeting less active accounts, most GNN-based models outperform the best non-GNN model. We highlight the best performing model for each column using the † symbol. If a GNN model performs better than or at least as well as XGBoost, we highlight it in bold.

TABLE V
THE AVERAGE SUSPECT VALUE FOR THE USERS OF DIFFERENT DEGREES OF ACTIVENESS

Percentile of active value	# spammers	Average suspect value
$[0\% - 10\%)$	222	16.487
$[10\% - 20\%)$	163	4.682
$[20\% - 30\%)$	86	3.577
$[30\% - 40\%)$	59	2.778
$[40\% - 50\%)$	57	1.931
$[50\% - 60\%)$	76	1.579
$[60\% - 70\%)$	112	1.126
$[70\% - 80\%)$	67	0.784
$[80\% - 90\%)$	51	0.511
$[90\% - 100\%]$	19	0.123
$[0\% - 100\%]$	912	5.899

2) Social-related features: The previous section illustrates that integrating social information helps identify less active spammers. This discovery led us to hypothesize that by designing social-related features, we could potentially assist non-GNN models in detecting less active cyberwarriors.

We introduce a new feature, the suspect value s_i, for each account i. The suspect value s_i is defined as the ratio of the number of times user i co-occurs with any spammer in an article to user i's activeness value, as expressed by Equation 1.

$$s_i = \frac{\sum_{\forall p \in \mathcal{A}_i} I(p \in \mathcal{P}^{(\text{spammer})})}{a_i}, \tag{1}$$

where a_i represents the activeness value of user i, $\mathcal{A}_i$ denotes the set of articles associated with user i (i.e., the set of articles in which user i has either posted or commented), $\mathcal{P}^{(\text{spammer})}$ returns the set of articles posted or commented on by spammers, and I denotes the indicator function, such that $I(x) = 1$ if x is true and 0 otherwise.

Table V presents the average suspect values for different ranges of activeness values. The results reveal a clear relationship between a user's activeness and their suspect value: users with lower activity levels tend to connect with more spammer accounts. This finding supports our earlier observation in Section IV-A that spammers exhibit less activity. Specifically, we find that inactive accounts tend to have more connections to spammers, which could indicate suspicious behavior.

Table VI reports the AUPRC scores of both non-GNN models and GNN-based models to detect cyberwarriors in different degrees of activeness, incorporating the suspect value as a feature. This social feature enhances the identification of cyberwarriors for both non-GNN and GNN-based models (referring to Table III and Table IV). Additionally, the suspect value feature proves particularly helpful in identifying spammers from the inactive user groups, as exemplified by LightGBM's AUPRC increasing from 0.49 to a remarkable 0.86 for the most inactive group of users.

TABLE VI
A COMPARISON OF VARIOUS MODELS (INCLUDING SOCIAL FEATURES) IN TERMS OF THE AUPRC SCORE. WE REPEAT EACH EXPERIMENT 10 TIMES AND REPORT THE MEAN ± STANDARD DEVIATION. WE HIGHLIGHT THE WINNER OF NON-GNN-BASED MODELS IN BOLD. WE HIGHLIGHT THE GNN-BASED MODELS IF THEY OUTPERFORM THE BEST NON-GNN-BASED MODELS.

Type	Model	[0%, 10%)	[10%, 20%)	[80%, 100%]	[0%, 100%]
Non-GNN-based models (including social features)	XGBoost	0.83 ± 0.01	$\mathbf{0.74} \pm 0.03$	$\mathbf{0.90} \pm 0.02$	$\mathbf{0.86} \pm 0.00$
	LightGBM	$\mathbf{0.86} \pm 0.02$	0.72 ± 0.05	0.88 ± 0.02	0.82 ± 0.00
	Random Forest	0.85 ± 0.01	0.56 ± 0.05	0.85 ± 0.02	0.79 ± 0.00
	Fully Connected	0.53 ± 0.07	0.51 ± 0.06	0.76 ± 0.05	0.64 ± 0.04
	ConvNet	0.43 ± 0.09	0.68 ± 0.07	0.83 ± 0.04	0.66 ± 0.06
	Soft Voting [22]	0.69 ± 0.00	0.56 ± 0.01	0.76 ± 0.01	0.72 ± 0.00
	Hard Voting [22]	0.67 ± 0.01	0.63 ± 0.02	0.70 ± 0.03	0.74 ± 0.01
	Stacking [22]	0.54 ± 0.02	0.56 ± 0.03	0.67 ± 0.01	0.69 ± 0.02
GNN-based models (including social features)	GCN	0.62 ± 0.08	0.52 ± 0.05	0.83 ± 0.08	0.69 ± 0.03
	TAGCN ($K = 1$)	0.79 ± 0.03	$\mathbf{0.97} \pm 0.05$	$\mathbf{0.99} \pm 0.04$	$\mathbf{0.92} \pm 0.01$
	TAGCN ($K = 2$)	0.82 ± 0.03	$\mathbf{0.98} \pm 0.02$	$\mathbf{0.99} \pm 0.03$	$\mathbf{0.93} \pm 0.02$
	TAGCN ($K = 3$)	0.85 ± 0.02	$\mathbf{0.98} \pm 0.03$	$\mathbf{0.98} \pm 0.01$	$\mathbf{0.94} \pm 0.01$
	GAT	0.73 ± 0.06	$\mathbf{0.91} \pm 0.06$	$\mathbf{0.92} \pm 0.07$	$\mathbf{0.87} \pm 0.05$

D. F1 scores When Claiming the Top-k Suspicious Users as cyberwarriors

After a model predicts the probability of an account being abnormal for each user, practical verification from administrators is still necessary. Therefore, a two-step procedure can be employed to determine suspicious accounts in practice. The procedure involves ranking all users based on their predicted suspiciousness using a prediction model, followed by manual examination of the top-k most suspicious accounts by administrators or guardians. The value of k is determined based on the available manpower, allowing for the verification of suspicious accounts with minimal labor costs.

To evaluate the effectiveness of the aforementioned two-step approach using different prediction models, we compute the $F1$-at-k ($F1@k$) scores for varying values of k. The $F1@k$ score is defined in Equation 2.

$$F1@k = 2 \times \frac{p@k \times r@k}{p@k + r@k}, \tag{2}$$

where $p@k$ and $r@k$ are precision-at-k and recall-at-k, defined by Equation 3 and Equation 4, respectively.

$$p@k = \frac{\text{number of abnormal accounts in top } k}{k} \tag{3}$$

$$r@k = \frac{\text{number of abnormal accounts in top } k}{\text{total number of abnormal accounts}} \tag{4}$$

The $F1@k$ score extends the standard $F1$ measure to evaluate a ranked list by considering the top-k predictions. It provides a comprehensive assessment by integrating both $p@k$ and $r@k$. The precision-at-k ($p@k$) measures the proportion of abnormal accounts among the top-k suspicious accounts, indicating the accuracy of the model's predictions within the top-k positions. On the other hand, recall-at-k ($r@k$) focuses on the completeness of predictions by evaluating how many abnormal accounts are included among the top-k positions. It indicates the model's ability to identify and retrieve relevant items from the entire set. The harmonic mean of $p@k$ and $r@k$ is used to compute the $F1@k$ score, ensuring that both precision and recall contribute to the final evaluation.

Table VII presents the $F1@k$ scores for various models at different values of k. The results demonstrate that LightGBM and XGBoost remain the top-performing models of non-GNN-based approaches. However, GNN models consistently outperform the best non-GNN models on $F1@k$ for different k values. Therefore, when employing the two-step human-machine cooperation strategy described above, utilizing GNN-based models with social features remains a favorable option.

V. DISCUSSION

This paper contributes to understanding spammers' activeness and the challenges associated with their detection. By examining a real dataset from a large forum, we have provided insights into the prevalence of inactive spammers, which were largely overlooked as previous studies primarily focused on active spammers. Our findings emphasize the importance of considering spammers' activeness and highlight the need for caution when applying existing detection models developed mainly for active spammers. The insights gained from this research may shed light on the broader landscape of spam detection and underscore the significance of adapting detection techniques to encompass both active and inactive spammers.

Although our primary focus in this study was on political spammers, it is worth noting that the methodology and approach presented can be extended to address other types of spammers, e.g., commercial spam. The underlying principles and techniques – incorporating social information into the model – can be readily applied to different domains, enabling the detection and mitigation of spam in various contexts. This versatility enhances the practical applicability of our research and provides a foundation for developing effective detection mechanisms in other domains related to spamming.

Future investigations could explore additional dimensions of spammer behavior, such as the temporal dynamics of their activities or the evolving strategies employed by different

TABLE VII
A COMPARISON OF VARIOUS MODELS (INCLUDING SOCIAL FEATURES) IN TERMS OF THE F1 SCORE. WE HIGHLIGHT THE WINNER OF NON-GNN-BASED MODELS IN BOLD. WE HIGHLIGHT A GNN-BASED MODEL IF ITS RESULT OUTPERFORMS THE BEST NON-GNN MODEL.

Type	Model	$k = 100$	$k = 200$	$k = 300$	$k = 400$
Non-GNN-based models (including social features)	LightGBM	0.6078	**0.7729**	0.6832	**0.5969**
	XGBoost	**0.6431**	0.7676	**0.6915**	0.5866
	Random Forest	0.6042	0.7598	0.6811	0.5849
	ConvNet	0.6289	0.7154	0.6336	0.5523
	FC	0.4382	0.6162	0.5880	0.5352
	Ensemble	0.1594	0.2325	0.2193	0.2096
	Soft Voting	0.1838	0.3148	0.4208	0.5092
	Hard Voting	0.1640	0.2842	0.3977	0.488
GNN-based models (including social features)	GATC	0.4382	0.6319	**0.6916**	**0.6038**
	GCN	**0.6573**	0.6632	0.5549	0.4700
	TAGCN ($K = 1$)	**0.6926**	**0.8564**	0.6873	0.5695
	TAGCN ($K = 2$)	**0.6997**	**0.8669**	**0.7122**	**0.5970**
	TAGCN ($K = 3$)	**0.7067**	**0.8721**	**0.7164**	**0.6072**

spammers. Such endeavors will contribute to a more comprehensive understanding of spamming phenomena and facilitate the development of robust and adaptive detection methods to counteract the ever-evolving landscape of spam.

REFERENCES

[1] R. DiResta, K. Shaffer, B. Ruppel, D. Sullivan, R. Matney, R. Fox, J. Albright, and B. Johnson, "The tactics & tropes of the internet research agency," 2019.

[2] M. Xiao, P. Mozur, and G. Beltran, "Buying influence: How china manipulates facebook and twitter," Dec 2021, [Online; accessed 6-May 2022]. [Online]. Available: https://www.nytimes.com/interactive/2021/12/20/technology/china-facebook-twitter-influence-manipulation.html

[3] X. Hu, J. Tang, and H. Liu, "Online social spammer detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, no. 1, 2014.

[4] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 35–47.

[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.

[6] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao, "Spammer behavior analysis and detection in user generated content on social networks," in *2012 IEEE 32nd International Conference on Distributed Computing Systems*. IEEE, 2012, pp. 305–314.

[7] Y. Lu, L. Zhang, Y. Xiao, and Y. Li, "Simultaneously detecting fake reviews and review spammers using factor graph model," in *Proceedings of the 5th annual ACM web science conference*, 2013, pp. 225–233.

[8] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI '13. AAAI Press, 2013, p. 2633–2639.

[9] I. Lin, "3 things we learned from the 936 banned chinese twitter accounts," Aug 2019, [Online; accessed 19-Sep 2022]. [Online]. Available: https://international.thenewslens.com/article/123912

[10] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, "Collabseer: a search engine for collaboration discovery," in *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, 2011, pp. 231–240.

[11] H.-H. Chen, P. Treeratpituk, P. Mitra, and C. L. Giles, "Csseer: an expert recommendation system based on citeseerx," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, 2013, pp. 381–382.

[12] H.-H. Chen and C. L. Giles, "Ascos++ an asymmetric similarity measure for weighted networks to address the problem of simrank," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 10, no. 2, pp. 1–26, 2015.

[13] H.-H. Chen, L. Gou, X. Zhang, and C. L. Giles, "Discovering missing links in networks using vertex similarity measures," in *Proceedings of the 27th annual ACM symposium on applied computing*, 2012, pp. 138–143.

[14] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[15] W. Fan, Y. Ma, Q. Li, Y. He, E. Zhao, J. Tang, and D. Yin, "Graph neural networks for social recommendation," in *The world wide web conference*, 2019, pp. 417–426.

[16] D. Y. Wu, T.-H. Lin, X.-R. Zhang, C.-P. Chen, J.-H. Chen, and H.-H. Chen, "Detecting inaccurate sensors on a large-scale sensor network using centralized and localized graph neural networks," *IEEE Sensors Journal*, vol. 23, no. 15, pp. 16 446–16 455, 2023.

[17] G. Wang, S. Xie, B. Liu, and S. Y. Philip, "Review graph based online store review spammer detection," in *2011 IEEE 11th international conference on data mining*. IEEE, 2011, pp. 1242–1247.

[18] Y. Yang, R. Yang, Y. Li, K. Cui, Z. Yang, Y. Wang, J. Xu, and H. Xie, "Rosgas: Adaptive social bot detection with reinforced self-supervised gnn architecture search," *ACM Transactions on the Web*, 2022.

[19] F. Shi, Y. Cao, Y. Shang, Y. Zhou, C. Zhou, and J. Wu, "H2-fdetector: a gnn-based fraud detector with homophilic and heterophilic connections," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1486–1494.

[20] M. Huang, Y. Liu, X. Ao, K. Li, J. Chi, J. Feng, H. Yang, and Q. He, "Auc-oriented graph neural network for fraud detection," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 1311–1321.

[21] Wikipedia contributors, "Ptt bulletin board system," 2022, [Online; accessed 6-May-2022]. [Online]. Available: https://en.wikipedia.org/wiki/PTT_Bulletin_Board_System

[22] N.-L. Nguyen, M.-H. Wang, and C.-R. Dow, "Learning to recognize sockpuppets in online political discussions," *IEEE Systems Journal*, 2021.

[23] K. Boyd, V. S. Costa, J. Davis, and C. D. Page, "Unachievable region in precision-recall space and its effect on empirical evaluation," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ser. ICML'12. Madison, WI, USA: Omnipress, 2012, p. 1619–1626.

[24] J. Cook and V. Ramadas, "When to consult precision-recall curves," *The Stata Journal*, vol. 20, no. 1, pp. 131–148, 2020.

[25] M. Welling and T. N. Kipf, "Semi-supervised classification with graph convolutional networks," in *J. International Conference on Learning Representations*, 2017.

[26] J. Du, S. Zhang, G. Wu, J. M. Moura, and S. Kar, "Topology adaptive graph convolutional networks," *arXiv preprint arXiv:1710.10370*, 2017.

[27] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.