Detecting and Characterizing SMS Spearphishing Attacks

Mingxuan Liu
liumx18@mails.tsinghua.edu.cn
Tsinghua University, Beijing National Research Center for Information Science and Technology

Yiming Zhang
zhangyim17@mails.tsinghua.edu.cn
Tsinghua University

Baojun Liu
lbj@tsinghua.edu.cn
Tsinghua University

Zhou Li
zhou.li@uci.edu
University of California Irvine

Haixin Duan
duanhx@tsinghua.edu.cn
Tsinghua University, QI-ANXIN Technology Research Institute

Donghong Sun
sundh105@tsinghua.edu.cn
Tsinghua University, Beijing National Research Center for Information Science and Technology

ABSTRACT

Although spearphishing is a well-known security issue and has been widely researched, it is still an evolving threat with emerging forms. In recent years, Short Message Service (SMS) has been revealed as a new distribution channel for spearphishing messages, which already has caused a serious impact in the real world, but has not yet attracted enough attention from the academic community. In this paper, we report the first systemic study to spotlight this emerging threat, SMS spearphishing attack. Through cooperating with a leading security vendor, we obtain 31M real-world spam messages that span three months. We design and implement a novel NLP-based detection algorithm, and uncover 90,801 spearphishing messages on the entire dataset. And then, a large-scale measurement was performed on the detected messages to reveal and understand the characteristics of SMS spearphishing attack. Our findings are multi-fold. We discover that SMS spearphishing has a significant negative impact on the real-world, and a large number of victims have been affected. And the distribution of active illicit types between spearphishing message and common spam is quite inconsistent. At the micro-level, to evade detection and increase the probability of success, adversary campaigns have evolved a set of sophisticated strategies. Our research highlights the impact of SMS spearphishing attack is prominent. We call on different community to work together to mitigate this emerging security threat.

1 INTRODUCTION

with fraudulent content To increase success rates, spearphishing attackers tend to collect and exploit as much personal information as possible from victims. Then they carefully construct customized and deceptive content to disguise as trustworthy entities. In contrast to bulk spamming, victims are more likely to be attracted, confused and then deceived by spearphishing attacks. Over the past decade, spearphishing has grown to become one of the most serious and influential security threats [15]. As reported by FBI [46], billions of dollars have been lost due to spearphishing attacks in recent years. Adversaries were also found to spread malware [7] as well as interfering in political elections [61] through spearphishing.

As a highly profitable attack, a wide variety of communication channels have been known to be abused by spearphishing, including email [25, 26], telephony [38, 59] and even social media platform [52, 55]. However, recently, it has been observed that *Short Message Service (SMS)* is also being used to spread spearphishing attacks [43, 66], which has caused serious financial losses and become an emerging serious security threat. As shown in the two examples of Figure 1, adversaries would embed leaked personal information of victims (shown in bold text) into short messages to attract their attention, and lure victims to click on URLs or call other contacts (shown as underlined text) for subsequent scams. Compared with other channels, the widespread usage of cellphone contacts increases the chances of attackers obtaining related personal information of victims and then defraud them. However, this threat has not received enough attention from the security community and our understanding of it is quite limited.

Flight Phishing Message Dear **Name**, your **Flight MU*****, from Fuzhou to Nanjing, has been cancelled. Please contact Eastern Airlines 0371-65****19 for refunding. [Eastern Airlines]

Covid-19 Related Scam Message Hi, **Name**. Register on fy18.cn and you can get recharge and masks. Come on, Wuhan!

Figure 1: Examples of SMS spearphishing attacks.

Prior work. Previous research has gained a suite of key features to identify spearphishing activities, such as automated account registration (in social platform) [55] and sender spoofing (in email) [11, 25, 26]. However, the format of SMS is much less rich than social platforms (e.g., limited text length, no hashtags) and emails (e.g., no attachments, no headers), thus complex clustering and NLP semantic analysis are not applicable. As for SMS analysis, existing works focus on detecting spam messages through template-based clustering [4, 13, 22], topic analysis [36] and sending behaviors [32]. However, these methods are all aimed at common spam and can not distinguish the "high-risk" spearphishing attacks from them. Therefore, detecting and characterizing SMS spearphishing attacks is by no means a trivial task.

ACSAC '21, December 6–10, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8579-4/21/12…$15.00
https://doi.org/10.1145/1122445.1122456

Research questions. In this paper, we report the first systematic study on SMS spearphishing attacks to answer a set of questions that are critical to understanding its security risks, including: *What are the characteristics of SMS spearphishing attacks? How many end-users witness spearphishing SMS? Have the adversary campaigns evolved into sophisticated strategies? And finally, how to mitigate this emerging threat?*

Our study. To the best of our knowledge, there is no publicly available dataset for SMS spearphishing analysis. We made this study possible by cooperating with 360 Mobile Safe, a leading security application with millions of monthly active users in China [10]. The application is developed for the android platform and provides spam detection functions. It collected 31.97 million fraudulent spam message detection logs in three months, and provided this dataset to us as the basis for our research. We then manually identified 1,196 spearphishing messages as ground-truth by inspecting 50,000 randomly selected messages. The labeling process was also assisted by senior security experts, which is the first and essential step in this study.

To propose a scalable detection scheme, we conducted an empirical study leveraging the above dataset and gained three key observations of spearphishing SMS, including *luring information*, *exploiting payloads* and *fixed syntactic*. Then a novel detection system was proposed based on the above observations with a precision of 96.16% on the labeled dataset. We then ran this system on the entire fraudulent message dataset, and found a total of 90,801 SMS spearphishing messages. We also developed a learning-based multi-classifier that can accurately classify a spearphishing message into nice business categories. Based on the above processing, we also conducted a comprehensive measurement study of the SMS spearphishing ecosystem, to recover and characterize the crime scene of this emerging security threat.

Measurement findings. Our discoveries of measurement are multi-faceted, and we highlight a few major findings here. First, as the realistic impact, our measurement results show that at least 24,346 victims were suffered from SMS spearphishing attacks during the three-month data collection period. In addition to the Names of victims, we found four other types of personal information, including"Flight Info", "License Plate", "Bank Card" and "ID Number" was also leaked and exploited to construct spearphishing messages.

At a macro-level, SMS spearphishing attacks behave unique features compared to spamming activities in other fields. Driven by profit, "Financial Scam" (40.86%) was the most active business in SMS spearphishing. We also observed two emerging categories of messages, "Lawsuit Scam" (27.11%) and "Fortune-telling Scam" (14.43%), which are novel businesses that have never been discussed in existing spamming analysis work. Furthermore, the working hours of attackers differ across various categories, e.g., Financial Scams tend to be active during weekdays and working hours, while Fortune-telling Scams mainly occur at night. Besides, for the infrastructures of this attack, we find that several well-known SMS gateways are extensively abused for distributing spearphishing messages due to their low cost and ease-of-use natures.

At the micro-level, we further explore the behaviors and strategies of attackers by grouping spearphishing messages into 11,475 campaigns. Based on the campaign-level analysis, we find several specific spearphishing businesses exhibit oligopolistic characteristics, i.e., a few campaigns dominate the major market share. We also discover that, to evade the detection and increase the probability of success, four strategies of spearphishing attackers are developed, including: (1) testing-sending; (2) progressive deception; (3) multi-semantic evasion; and (4) global affair integration. Our research is the first to reveal the above strategies, which could help the community to better understand the SMS spearphishing attacks and provide assistance in the mitigation.

Overall, our research not only spotlights the emerging security threats to SMS spearphishing attacks, but also provides in-depth insights to help understand and mitigate the threat.

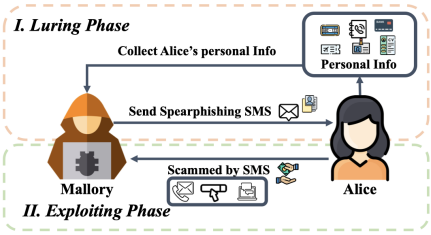


Figure 2: The threat model of SMS Spearphishing Attack.

2 BACKGROUND

Spam SMS typically refers to unwanted or unsolicited messages received by mobile phone users via Short Messaging Service (SMS). Previous researches demonstrated that, Spam SMS could be generated through various channels, including fake base stations [33, 47, 73] or SMS Gateways [48, 49], and are primarily associated with illegal promotion or malware distribution [14, 22, 40, 70].

Although several works have been devoted to detecting and analyzing the spam SMS ecosystem [36, 48, 49, 73], adversaries are also constantly evolving and new threats are emerging in this field. This study focuses on one of the most insidious threats, SMS-based spearphishing attacks. The adversary crafts a *specifically targeted* short messages with *fraudulent content* to trick the victim into performing a dangerous action. This selective targeting and motivation differentiate SMS spearphishing attacks (our research) from common Spam SMS, which is generally sent in an undifferentiated bulk manner. One may also refer this attack to spearphishing in emails, while it is worth noting one fundamental difference between them: the preparatory information required for attackers to construct targeted fraudulent content. Specifically, an email address naturally implies some information of a "username", which could be used directly to construct deceptive content (e.g., a fraudulent email starting with "Hi, username"). However, the same facility could not be offered by phone-number. Therefore, a spearphishing email attacker can carry out a targeted scam simply by getting the victim's email address. While for speraphishing SMS attackers, the collection of meaningful PIIs (personal identifier information) besides phone-numbers, such as the victims' Name and ID number, becomes an extra prerequisite. The personal information may require additional stealing operations, or be obtained from existing leaked database.

Threat Model. Based on the above analysis, we could summarise the threat model of spearphishing SMS as Figure 2. Assume that Mallory is the attacker, who tries to fool a targeted victim, Alice,

through targeted deceptive SMS content, and then exploits the trust to induce Alice to perform dangerous actions. The attack process could be divided into two phases, "luring" and "exploiting".

During the *luring* phase, Mallory is required to gain Alice's trust, i.e., preparing to frame sufficiently deceptive fraudulent messages. To achieve this, Mallory would first extensively collect Alice's personal information, such as the Phone-Name pair(and possibly other PIIs, e.g. ID card), and then disguise Alice as a trusted entity by crafting and sending fraud SMS messages embedded with Alice's PIIs.

In addition to constructing targeted content to gain Alice's trust, Mallory would also embed malicious payloads into the SMS content to trick Alice into executing them for tangible profit, which is termed as the *exploiting* phase. Specifically, this process happens after Alice is scammed by the carefully crafted fraudulent content. Then, Alice will execute various types of embedded payloads in the text, i.e., several follow-up contacts that trigger subsequent fraudulent activity, including: (1) clicking on URLs that distribute malware or steal credentials, and (2) triggering out-of-band actions (e.g., wiring money or making phone calls).

In other words, we focus on attacks where attackers harvest large-scale victims' personal information and craft fraudulent spearphishing short messages to masquerading as trusted entities.

Research goal. To perform the first exploration of the spearphishing SMS ecosystem, one prerequisite is obtaining a real-world dataset as the research basis. Perhaps the most ideal way is that, we as the researchers to design and deploy one detector operating on real-world clients that distinguish spearphishing SMS from legitimate ones directly. However, it is not considered in this work for ethical reasons. Directly detecting spearphishing requires the researchers to monitor, read, and process all the SMS data received by real users. It meanwhile inevitably exposes the researchers to legitimate SMS messages, which may contain sensitive information, and thus poses serious privacy risks.

Therefore, we chose to "obtain a sound dataset" rather than "design a perfect detector". As the message content of spearphishing SMS is fraudulent, which also falls into the general scope of spam SMS, we decided to work with a mobile security vendor. The security application of this vendor has deployed the spam SMS detection module on mobile clients and accumulated a real-world spam SMS dataset with well scale and coverage. We then collect the research dataset by further detecting spearphishing out of the already discovered spam pool. We acknowledge that, the attacks we identified may be limited by the view scope of the security application. However, it is a practical solution to balance the ethics and usability, which is adequate to give a first glimpse of the spearphishing SMS ecosystem. Considering currently known spearphishing SMS are basically user-reported cases, our automated approach will serve as a valuable improvement. Besides, the findings of this work are also expected to help design a more integrated detection system in the future.

3 DATASETS

In collaboration with 360 Mobile Safe [1], we performed a data-driven study to explore spearphishing attacks through SMS. In this section, we will elaborate on the details and ethical considerations of data collection.

3.1 Data Collection Process

360 Mobile Safe is a mobile security application in China. It is available for all versions of the Android platform, and mainly serves Chinese users [41]. Currently, it is promoting on most popular app stores, but not on Google Play as Google is blocked in China.

To prevent end-users from being harassed by spamming, 360 Mobile Safe provides the functionality to detect and filter spam messages. More specifically, once a message is received, it extracts the message content and sender information, and then a local SVM classifier and an online deep learning system work together to identify spam behaviors based on the information collected. If a message is detected as spam, it would be transferred into Spam Inbox, and the user will receive a pop-up notification. It should be noticed that the end-user still has the right to manually recover the message, if one message "looks safe".

Furthermore, to help the software improve detection algorithm and capture spam campaigns for law enforcement agencies, a spam detection log would be generated and uploaded to cloud servers for further security analysis. As for the detection log, it contains not only the message content and sender information, but also timestamp, city location, and hashed International Mobile Equipment Identity (IMEI) and International Mobile Subscriber Identity (IMSI) by the SHA256 algorithm (to protect user privacy).

For this work, our industry partner provided us with all the spam detection logs, except for spam classified as commercial marketing promotion with two considerations. The first one is to protect user privacy, as the percentage of commercial promotion is significantly high (>99% of all spam logs), it may leak sensitive information of users, such as browsing habits. Second, SMS spearphishing attacks are less likely to appear in marketing messages. However, other fraudulent messages are commonly used to launch malicious activities, such as phishing and illegal promotion. Therefore, their related data can be shared with researchers for security analysis.

Totally, we collected three months spam detection log from December 28, 2019 to March 25, 2020, containing **31,956,437 fraudulent messages**, and covering all provinces in China.

Limitation of Dataset. Although we tried to make this study as comprehensive as possible, there are still some limitations here. First, our dataset is collected from a mobile application in China, so it may have a geographical bias due to the user distribution. However, the long-term large user base of our industry partner and millions of monthly active users of 360 Mobile Safe make a very comprehensive national coverage. Therefore, the collected dataset could be comprehensive enough for a country-level study. Second, to protect user's sensitive information, we only collect the spam detection logs which are not identified as commercial marketing promotion. SMS spearphishing attacks are less likely to appear in marketing messages, based on the studies on the nature of spearphishing attacks via other channels [11, 25, 26, 55]. So the SMS spearphishing attacks we detected are representative. Dataset collection in previous works only focuses on the common spam SMS [48] and spearphishing attacks via other channels, e.g. email [11, 26]. To the best of our knowledge, our dataset is the first one from which we can detect considerable SMS spearphishing attacks.

3.2 Ethical Considerations

The nature of our research, i.e., detecting SMS spearphishing attacks embedded with personally identifiable information (PII), dictates the challenges we must face when dealing with sensitive datasets. As Institutional Review Board (IRB) has rarely been established in Chinese research institutions yet, we were unable to obtain an IRB approval. Nevertheless, we take the utmost effort to complement the review function of IRB. Specifically, the entire study was conducted during the researchers' internship within that company. All their operations, including data collection, data analysis and data storage, complied with the ethical requirements in the cooperation agreement. And the entire data processing steps were supervised by the company's legal committee. In addition, we carefully adhere to ethical guidelines for cybersecurity research, including the recommendations from Partridge and Allman [3], and the Menlo Report [16]. Below we discuss the ethical considerations in detail before presenting our methodology.

Data Collection. (1) The collection of spam detection logs is dominated by our industrial partner. The entire process has proven to be in strict compliance with the data privacy policies of the 360 Mobile Safe legal committee, and is subject to their oversight; (2) When installing the security application, users receive a consent [51] form that details what types of data would be collected, how their privacy would be protected and the usages of the collected data. Specifically, it also clearly states that the data could be provided to research institutions for academic usages. Meanwhile, one necessary condition is, all published academic results must ensure that no sensitive information would be released. In this work, we take adequate anonymization effort and avoid making any potentially sensitive analysis. We believe it satisfies the condition of academic data usage as stated in the user consent. Besides, users are also informed of the benefits and potential risks of turning on spam detection, and then voluntarily decide whether to join in and have the right to opt-out at any time.

Data Analysis. During the data analysis process, we work closely with professional lawyers to ensure that each step is legal. We also tried our best to balance the beneficence of experiments and the potential risks. (1) We took much effort with our industrial partner on dataset anonymization. First, all the device-related identifiers, including the IMSI(unique identifier of SIM Card) and IMEI (unique identifier of Mobile Equipment) of mobile phone users, were hashed before being provided to us. Then, through a manual inspection on a small dataset in empirical study, we found several types of victims' PIIs could be embedded in the message content, such as victims' Name, ID Numbers and Flight Information. Then, regular expressions were built to detect and replace these PIIs with hashes by scripts (see Section 4.1 for details). It ensures that the researchers would not be exposed to sensitive information as much as possible; (2) We signed a cooperation agreement with the security vendor to ensure the data processing is completed on the company's virtual environment and all the data is kept confidential; (3) We also double-checked the measurement findings published in this paper, to ensure that no personal information was inadvertently disclosed.

Data Storage. All the detection logs are stored on confidential servers within the industrial company with security reinforcements. Investigators accessed the data as interns. All data are not allowed to be copied to external networks.

In summary, we have employed a set of best practices to mitigate potential ethical concerns. And we believe the beneficence of the first spotlight on SMS spearphishing attack and further understanding of this threat outweigh its potential ethical risks.

4 METHODOLOGY

In this section, we first present an empirical study of manually labeling and inspecting ground-truth dataset as a guideline. Then, leveraging three insights gained from the empirical study, we propose a detection system that is able to detect SMS spearphishing attack from fraudulent messages. An overview of the system architecture, implementation details and evaluation results are elaborated in the following subsections.

4.1 Empirical Study

Ground-truth dataset. To the best of our knowledge, there is no public dataset available for SMS spearphishing attack. Therefore, we bootstrap our study by inspecting a huge number of fraudulent SMS collected by 360 Mobile Safe, and manually label spearphishing messages to create the ground-truth dataset.

We randomly selected 50,000 fraudulent messages from the entire dataset assembled a labeling team of two investigators. To establish consensus among members, we first sampled a set of 5,000 messages (10%) for the investigators to label independently, and reached an agreement score of 89.30%. Following, a senior mobile security expert was invited to review and discuss the inconsistencies results with our team. During this process, a set of empirical guidelines were summarized to distinguish SMS spearphishing attacks from regular spam messages. After review, all conflicts in the first round labeling were resolved, and no inconsistent results appeared for the labeling of the remaining 90% messages.

In total, we labeled *1,196 messages* (2.39% of 50,000) as spearphishing and regarded this dataset as ground-truth. To factorize business types of spearphishing attacks, we also provide an empirical multi-classification of messages, which will be described in Sec 4.3.

Key observations. Traditional detection mechanisms of spearphishing attacks (e.g., email) are effective when they behave spoofing or 'phishy' emotion [11, 69] identified by Natural Language Processing (NLP) tools. However, in our case, those approaches are not effective. The most significant challenge is that the length of SMS text is too short, which may lead to unpredictable errors in conventional NLP topic modeling and sentiment analysis [53, 62]. As such, new features are needed to distinguish spearphishing SMS attacks in the context of spamming. We discover three key observations through empirical analysis of ground-truth, which can help to build the detection system, as elaborated below.

- *"Luring": Personal information of victims.* To make victims feel trustworthy or familiar, we observed that all spearphishing messages were customized by the victim's **Name** (or Last Name). In several special scenarios, Flight Information, Plate Number, Bank Card Number and ID Card Number were even included to enhance the allure.
- *"Exploiting": Out-of-band contacts of attackers.* From the view of the attacker, it is unprofitable to simply spread unsolicited information. An adversary must embed at least one follow-up contact (e.g., click on URLs, contact with social accounts)

to carry out subsequent fraudulent activity. One may argue the need to embed contact in the message since the sender of the message is a natural contact to call back. However, it is more costly for attackers to scam by calling back directly [59] as it requires maintaining a fixed contact device and serving real-time human interaction on it all the time.

- *"Syntactic": Syntactic relationship of personal pronouns.* However, the above two features alone cannot accurately distinguish SMS spearphishing attacks from illegal promotions. Since some attackers may also embed their Names in messages, as the bottom case shown in Figure 4. Therefore, we need not only to identify the embedded personal information, but also to determine whether it belongs to the victim. Fortunately, we observe that the texts of spearphishing messages are often crafted in the tone of a conversation between the attacker and the victim, while the attacker is always the initiator. With the help of syntactic analysis in NLP, we find a difference between attacker and victim in the syntactic structures of their pronouns.

Anonymization. By manually inspecting the raw data of 1,196 labeled spearphishing SMS, we found 5 types of PIIs could be embedded in the message content, as shown in Table 4. For privacy reasons, these PIIs should also be anonymized, and we take this anonymization as a necessary step in the data processing workflow of this work (as part of **Step II Entity Recognition**, see Section 4.2). Specifically, Name of the victim could be identified by Name Entity Recognition (NER) and then be replaced by hashes directly. Although the other four types of PII are not typical "Entity" in Natural Language Processing (NLP), they have certain string formats which could be identified by regular matching. We manually built the specific regular expression for each type of the four PIIs, as examples presented in the last column of Table 4. Our industry partner then helped detect and replace all these PIIs with hashes on the entire dataset, using scripts we built based on NER and the above regular expressions. Afterwards, the sufficiently anonymized dataset was re-provided to us for subsequent data processing and analysis.

4.2 System Design and Implementation

Design Overview. Inspired by the three observations gained from the empirical study, we are able to design a detection system. This system aims to accurately identify SMS spearphishing attacks through collaborating with a mobile security application that detects unsolicited messages at the client-side.

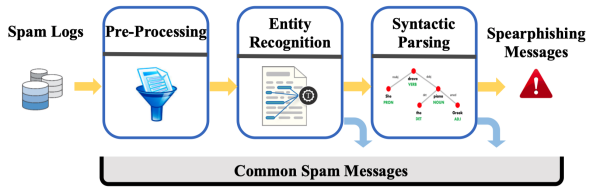


Figure 3: Overview of the detection system.

Figure 3 abstracts the system workflow. As the first step, the system extracts necessary fields from spam detection logs, and tries to process obfuscated text. The pre-processed text goes to the *Entity Recognition* module, which attempts to find customized personal information and contacts. Subsequently, the text with labeled entities is forwarded to the *Syntactic Parsing* module, which distinguishes the attribution of extracted personal information. If the personal information belongs to the victim, the message will be marked as a spearphishing message. Otherwise, the message will be considered as regular promotion (labeled as "common spam").

Step I: Data Pre-Processing. During the empirical study, we observe that attackers often utilize text obfuscation to bypass detection mechanisms. As NLP techniques are generally suitable for well-written text, the obfuscated and ungrammatical content, i.e., adversarial text, may significantly reduce the effectiveness and reliability of NLP tools [27, 30, 75]. Therefore, it is necessary to "sanitize" the text content of fraudulent messages before forwarding them to the next module.

Specifically, as the first step, we summarized the most popular obfuscation methods by examining ground-truth dataset: mixed text with special characters. For example, punctuation from different languages could be mixed, especially between Chinese and English, e.g. the dot "." in domain names could be replaced by "。" or "．", and the ":" in a URL could be replaced by "：". To solve this issue, we removed all redundant spaces and special characters by comparing them with the public character list [24]. Also, digits could be replaced by characters that are visually similar to them, e.g. digit "1" and letter "l", and digit "0" and letter "O". Inspired by previous works [31, 75] on adversarial text, we replaced these characters with common morph combinations.

In addition, we also observed that illegal promoters create a jargon term ("black keywords") to disguise transactions. Black keywords are often unfriendly to outsiders, distorting the original meaning of common terms or tweaking other black keywords. For example, "微信" (Wechat) can be replaced by "徽信" which looks similar but is not an existing word in Chinese. Since existing name entity recognition systems are most domain-specific, it is difficult to properly adapt and label these specialized terms. To address this issue, we note that a previous research collected and built a list of jargon words for the Chinese underground economy [68]. In order to reduce the errors introduced by these "black keywords", we extend Yang's jargon list based on the observation of empirical study, resulting in 4,718 jargon words. Then we replace the jargon terms in the list with a fixed word "JARGON".

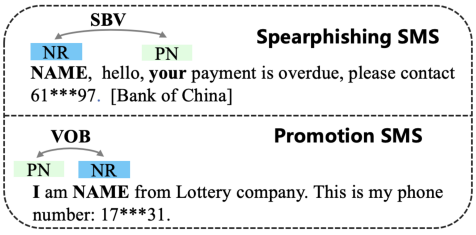
Step II: Entity Recognition. Recall our observation obtained from the empirical study that SMS spearphishing attacks must include *luring* information and *exploiting* payloads. In the following procedure, we attempt to identify the customized personal information and out-of-band contacts contained in suspicious messages by entity recognition.

Based on the examination of ground-truth, we treat five types of "victim's personal information", including Name, ID Number, Flight Information, License Plate Number and Bank Card Number, as PIIs. These PIIs need to be both identified as the *luring* information for spearphishing detection and anonymized for privacy concerns. As discussed in Section 3.2 and 4.1, we recognize human names by NER through a open-source tool, HanLP 2.0 [24], which has a well adaptability to Chinese text. And we built regular expressions to extract the other four PIIs. The recognization and anonymization process was automated by scripts we built. Using the scripts, our industrial partner helped, i.e., identifying and then replacing all

the PIIs with their hashes. We believe this step could minimize potential harm and ensures that this research is ethically sound.

In addition, we also extracted the embedded contacts of fraudsters in the messages for follow-up communications with victims. As mentioned in Section 4.1, embedded contacts are also a key feature of a spearphishing message. We design and implement regular expressions to detect and extract several of the most commonly used contacts, such as Phone Number (Cellphone, Phone and Hotline), URL/Domain, Wechat and QQ (two of the largest social platforms in China).

To sum up, suspicious messages containing both the victim's personal information and attacker's contact would be forwarded to the next module.



* NR: Name. PN: Pronoun. SBV: Subject-Verb. VOB: Verb-Object.

Figure 4: Examples of syntactic parsing on spearphishing and promotion messages[1].

Step III: Syntactic Parsing. Syntactic analysis is able to examine the syntax dependency in a given sentence, which is one of the most important technologies in the NLP research field. In this module, we utilize syntactic parsing to examine the ownership of the extracted PII entities. As discussed in Section 4.1, one of our key observations is that, a Name belonging to the victim differs in syntax features from the one belonging to the attacker. Since SMS spearphishing texts are often sent in a tone of the conversation between the attacker and the victim, in which attacker is the initiator and victim is the recipient. From the perspective of syntactic analysis, we found that names of victims commonly meet the following two relationships with its personal pronouns: Subject-verb relationship and Attributive modification. As supporting evidence, Figure 4 shows an example of a spearphishing message with the Subject-verb relationship. By contrast, a common promotional message doesn't embed the victim's PII. The self-introducing phrase shows the Verb-object relationship with its personal pronouns, which is different from the relationship of victim's.

As a result, after extracting personal information, we leverage the syntactic parsing function of HanLP [24] to check the syntax dependency. Only the SMS message that matches the above two kinds of relationships would be regarded as spearphishing attack. Otherwise, the messages are still considered as regular fraudulent spamming.

4.3 Classifying Business Categories

To perform a large-scale measurement study and understand the ecosystem of SMS spearphishing attack, we need to classify the business types of spearphishing content. Due to the lack of public labeled datasets, we first created a self-labeled dataset, and then trained a machine-learning model to construct the multi-classifier.

[1]The examples are translated from Chinese messages.

Table 1: Categories of spam messages.

Category	Volume in Labeled dataset	Volume in Common SMS	Volume in Spearphishing SMS
Financial Scam	1,943	2,583,017	24,668
Lawsuit Scam	2,206	2,452,277	13,124
Social Scam	271	672,853	2,608
Employment Scam	1,019	1,244,083	10,620
Insurance Scam	86	93,670	739
Fortune-telling Scam	1,363	1,226,959	15,500
Gambling Phishing	1,918	15,598,262	16,319
Promotional Spam	1,263	5,775,055	6,201
Other	330	2,310,261	1,022
ALL	10,399	31,956,437	90,801

Table 2: Detected information of spearphishing spam SMS.

Victim's Information			Spammer's Contacts		
Entity	**Record**	**Content**	**Entity**	**Record**	**Content**
Name	90,801	71,655	URL	58,968	45,935
Flight	883	10	CellPhone	13,158	11,169
License Plate	571	536	Hotline	3,922	2,627
ID Card	17	13	Phone	3,419	1,882
Bank Card	1	1	QQ	11,959	9,599
			WeChat	5,215	4,359

Labeled Dataset. At this step, we randomly selected 15,000 samples from the entire dataset, and 10,399 messages are kept after data deduplication. Two members labeled them independently according to nine pre-defined categories, which were delineated with reference to several published technical reports [44, 50] and papers [48]. In the first round, the agreement score is 98.24%, and then we also discussed with a senior security expert to solve 183 conflicts. After review and discussion, we gave each conflict a unanimous agreement label. The volumes of each category are shown in Table 1, with the Lawsuit Scam accounting for the most in the labeled dataset.

Multi-classifier Models. In our study, 10,399 messages with consistent label were considered as the labeled dataset (for business classifier), under nine categories. We tried two popular methods of word embedding to get the vector representation, including Word2Vec [29] and TF-IDF[58]. Then we applied five popular machine learning models for text classification [28]. Leveraging our labeled dataset, Table 5 shows the performance of all text classification models with different embedding ways.

In the end, we find that the combination of Word2Vec and logistic regression model[21] performs the best (average F1-score 93.41%). Therefore, we employ it to categorize all other messages.

4.4 Evaluation Results

We implemented the detection system on the entire collected dataset and detected 90,801 (71,655 deduplicated message content) spearphishing messages. The detailed detection results are shown in Table 2, and representative examples for each category of spearphishing messages could be seen in Appendix **??**. Here, we discuss the evaluation results of our detection system and category classifier.

Effectiveness of Detection System. As for the ground-truth dataset (1,196 messages), 937 of them were correctly detected, with 36 false positives (precision 96.16%) and 292 false negatives (recall 75.33%). We randomly sampled 200 spearphishing messages and manually checked, with only 8 false positives. According to other works' evaluation [25, 73], the precision in the sample set, 96%,

combined with the precision in the ground-truth dataset, can represent the precision on the whole detected spearphishing messages, due to the randomness of sampling.

We manually checked the detection errors and found that the low recall rate is mainly limited by the performance of the entity recognition algorithm. Due to the domain-specific in spam content and the specificity of Chinese language, it is known that open-source NLP tools are difficult to achieve the desired performance[34, 74]. And until now, there have been no sophisticated solutions to this problem. We acknowledge that the recall rate of our detection system is not perfect, which means that our detection results are only the lower bound of actual SMS spearphishing attacks.

Category Classification Result. For the multi-classifier, the precision is 93.46% and recall is 93.47% on the labeled dataset. From the confusion matrix, we find most categories could be classified accurately with precision over 86%, except for "Other" and "Promotional Spam". Manual inspection of these two categories showed that, the main reason for misclassification was that the length of several messages is too short to extract semantics. However, these two categories are relatively least malicious among all spearphishing businesses, that we rarely focus on in the subsequent measurement analysis. Thus, we consider that the performance provides reliable results to support our measurement findings.

5 MEASUREMENT

Our detection system reports 90,801 (0.285% of total) spearphishing messages on the three-month dataset. In this section, we empirically analyze the behavior of spearphishing attackers based on the detected messages, including their sending characteristics (business categories, time and geographical characteristics), infrastructures (distribution channels and out-of-band contacts), real-world impact, and the personal information (in hash format) of victims. We also group the detected messages into spearphishing campaigns, and discussed their active properties and attacking strategies.

5.1 Characteristics

In this section, we will examine the spearphishing attacks from a macro perspective, including its business categories, sending behaviors and infrastructures (sending channels and follow-up contacts) that attackers utilize.

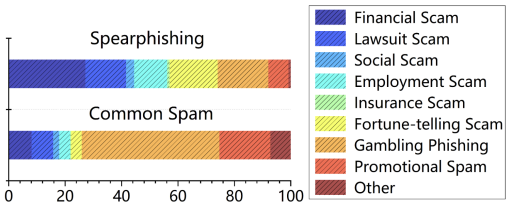


Figure 5: Comparison of the category distributions.

Spearphishing Categories. As shown in Figure 5, the proportion of active categories in spearphishing differs from common spam significantly. For common spam, Gambling Phishing, an illegal service which has developed into a mature underground industry in China [67], accounts for the majority(48.90%), followed by Promotional Spam(14.64%). However, Financial Scam (40.86%) is the most active business of spearphishing, in which the attackers deceive users by posing as reputable banks or financial companies and offer loan services. Based on our dataset, 8 banks were found maliciously disguised, including Bank of China and China Construction Bank. One example is given in Table 3, in which the attackers disguised as the Bank of China (BOC) claiming the victim met repayment troubles while leaving a personal mobile phone (instead of the official BOC hotline) as follow-up contacts. Besides, we also observed 6 loan apps sending spearphishing messages, claiming the loan was overdue thereby scaring the victim into clicking on the malicious shorten-URL (usually pointing to a phishing website) embedded in the messages.

Compared with previous studies of common spam [48, 49, 73], we observe that two new businesses are particularly active in spearphishing SMS ecosystem, Fortune-telling (27.11% of all) and Lawsuit Scam (14.43% of all). Fortune-telling is a novel scam that has never been discussed previously. Attackers usually claim that they can accurately predict the fate of victims, and provide the details through the embedded links. While as the example of this category shown in Table 3, the links usually indeed pointing to phishing websites. Lawsuit Scams are where the attackers claim that they would prosecute the victims for violating some legal provision, to threaten the victims into contacting them with left contacts, as the example shown in Table 3. Although the fraudulent techniques of the above two businesses are different, with Fortune-telling relying on luring the victims and Lawsuit threatening them, in both scenarios, the inclusion of victim's personal information is always the crucial step to make the scams more deceptive.

Sending Characteristics of Attackers. As mentioned in Section 3, the logs we obtained contain the time and geo-location information of the victim receiving the SMS. Thus, we could study the attackers' message sending behavior by examining the spatio-temporal distributions.

Based on the distribution of receiving time, we find the "working patterns" of attackers differ across spearphishing categories. For example, Financial Scam attackers tend to send messages intensively during weekdays and working hours. The reason may be that, Financial Scams are largely engaged in spoofing well-known financial institutions such as banks. As these institutions usually send messages during working time, attackers also mimic the same working pattern to make the disguise more realistic. In contrast, Fortune-telling Scam messages were mainly sent at night (over 87.67% were sent between 18:00 and 21:00). In this case, attackers were essentially marketing fortune prediction services, so they tend to operate at leisure time, leaving sufficient space for victims to read the information and purchase services.

The geographical distribution of spearphishing victims is shown in Figure 6, which is (not surprisingly) roughly proportional to the regional population distribution [8]. In particular, Guangdong receives a significantly higher volume of spearphishing messages, the vast majority of which are Gambling Phishing. This is determined by its unique geo-location: Guangdong is adjacent to Macau, the only region in China where gambling services are legal. Furthermore, although several less economically developed regions (e.g., Northwest China) received a lower absolute number of spearphishing messages due to their smaller populations. If we calculate the ratio of spearphishing to common spam in each region, i.e., the "spearphishing rate", their rankings are quite high. For example, Tibet, which has the lowest GDP in China (in 2019 [42]), ranks

Table 3: Representative spearphishing messages of each Category.

Category	Example of message content
Financial Scam	Dear Mr/Mrs. **NAME**, your BOC credit card has been suspended due to an overdue payment. Please contact 86137****765.
Lawsuit Scam	[CMB] **NAME**, your credit card has been deemed to be seriously overdue. We will formally prosecute you after 24 hours! Please contact: 86239****7834
Social Scam	**NAME**, how are you recently? I changed WeChat. Please add my new WECHAT 266****491.
Employment Scam	**NAME**, hello! Your resume in 5*.com investment has been accepted. Please contact the QQ: 33****471.
Insurance Scam	[China Life Insurance] Dear **NAME**, our company has issued a commercial insurance policy for your car, **PLATE**. Visit http://***.cc/bX8Vg for details
Fortune-telling Scam	[Lingji Culture] **NAME**. Full analysis of fortune in next ten years is coming! See the future and prevent bad luck: https://s.k****a.cn/dbgc8
Gambling Phishing	Hi, **NAME**. Yabo Sports join hands with Wuhan, register on f**8.cn and you can get recharge and get masks, come on, Wu Han!
Promotional Spam	[Tantan application] **NAME**, someone loves you, do you want to accept? It is only 3 Km away from you! Click tan****pp.com.
Other	Mr. **NAME**, do you want to purchase brand products at a discount? Please contact QQ: 324 *** 558.

only 31/34 in the absolute number of spearphishing SMS received, but has the highest "spearphishing rate" (0.41%). This suggests that attackers would tend to target victims in less-developed areas, possibly because people in these areas are relatively less educated, thus making it easier for scams to succeed.

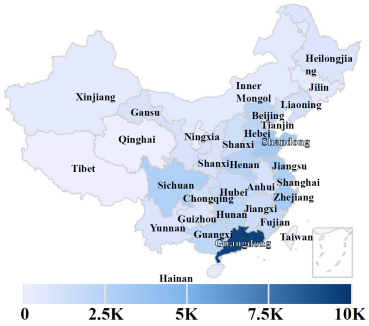


Figure 6: Spearphishing victims distribution.

5.2 Infrastructure

Distribution channels. We first focus on the channels through which the attackers distribute spearphishing messages by examining the sender numbers. To better characterize the sender numbers, we crawled a list of popular hotlines [12], area codes [19] and ISP numbers [60] as the reference. Based on our dataset, a total of 26,129 sending numbers are observed, of which 13.97% came from outside of China (possibly to reduce legal risks). Further inspection of these overseas senders reveals that, the vast majority of them were engaged in Gambling Phishing (72.78%) and mainly located in the Philippines (44.06%). The remaining senders inside China consist of three main sources: 106 SMS platforms (43.65%), cellphones (41.78%) and hotlines (0.61%). The 106 SMS platform is a special case in China, which refers to the SMS gateway provided by Chinese Internet Service Providers such as China Mobile [39] and China Telecom [54], and enables bulk SMS sending functions at low prices [37]. Messages sent from this platform would display with a virtual sender number starting with "106". Unfortunately, its low-cost and easy-to-use nature has also attracted the attention of underground industries. As reported in [45], in 2018, up to 92% of the spam bulk messages were distributed from 106 platforms. Our study further confirms that, the 106 SMS platform has also become the "workhorse" of spearphishing SMS attacks. Besides, we find that different spearphishing categories tend to use different sending facilities. For example, over half of Gambling Phishing messages come from foreign senders, while almost 80% of Financial Scams utilize the 106 SMS platform.

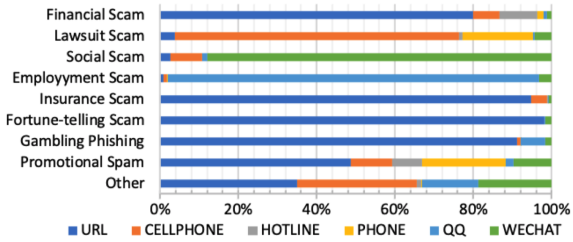


Figure 7: Percentage of contacts used by category.

Out-of-band contacts. We also examined the contacts left in messages as the follow-up communication methods between spearphishing attackers and victims. As shown in Table 2, URLs are the most common ways, typically used to promote websites or distribute malicious APKs. Attackers also leave phone numbers (Cellphone, Hotline, Phone), or provide social accounts (QQ, WeChat) for follow-up communication. Moreover, the nature of different spearphishing businesses can affect the attacker's propensity of contacts selection, as shown in Figure 7. Employment Scam relies extensively on social platforms to post recruitment tasks (see Example in Table 3), thus they prefer to use QQ as the contacts, accounting for 94.93%. WeChat is one of the most popular social platforms in China, and Social Scam often deceives victims to add WeChat (87.92%) to perform subsequent scam activities. In addition, since Lawsuit attackers usually require human interactions to complete the scam, they are more likely to take telephones (72.64%) as follow-up contacts.

5.3 Campaign Analysis

In order to gain deeper insights into SMS spear spamming activities, we further explore characteristics from a higher perspective, i.e., by grouping the detected spearphishing messages into spam campaigns. Specifically, we treat the messages embedded with **the same contact** or from **the same content template**, while with the Levenshtein Distance [71] less than 5, as being sent by the same campaign. The threshold of Levenshtein Distance is an empirical value referred to previous work [73]. And we conducted manual inspections to confirm it could work effectively on our dataset as well. A total of 11,475 campaigns were reported in this way. Active properties and several interesting strategies of campaigns are described as follows.

First, the scale of spearphishing campaigns exhibits a long-tail distribution, with the top 100 campaigns accounting for 48.78% of all spearphishing messages, and the top 1,000 campaigns accounting for 73.19%. The largest campaign with 14,561 messages (16.04% of all) is engaged in Fortune-telling Scams. Further inspection revealed that this campaign was bursty in nature, with all messages sent intensively during a 23-day period in March 2020. It is also the

most influential campaign, with a total of 9,667 impacted victim devices spread across 34 provinces in China. As for the rest, 71 campaigns affected more than 50 victims, and the average number of affected victims is 5 across all the campaigns. Further investigation into the relationship between campaigns and the affected victims revealed that, leaked personal information could be obtained by multiple criminal groups, exposing the victims to a variety of scam threats. Concretely, we found 3,155 devices (7.44% of all) received spearphishing messages from two or more campaigns, with the most unfortunate one affected by up to 18 campaigns involving businesses of Gambling Phishing, Social Scam and Promotional Scam.

We then investigate how different spearphishing categories are distributed among campaigns and find that, the business of Financial Scam, Insurance Scam and Fortune-telling Scam all exhibit oligopolistic characters, i.e., a few campaigns hold the majority of the market share (i.e., the volume of spearphishing messages). Specifically, the largest campaigns of Fortune-telling Scam and Insurance Scam both hold more than 50% of their messages. For Financial Scam, over 50% of the messages were split between campaigns of the top 6. Furthermore, although 97.04% of campaigns are engaged in only one type of spear scam, the few remaining ones that undertake multiple categories present interesting business models of spearphishing. Figure 8 shows that 11 of the top one hundred campaigns engaged in more than one category. For example, we found campaigns operating Financial Scam tend to also conduct Lawsuit Scam (accounting for 2.96% of campaigns engaging in multiple businesses). Further inspection of the message content reveals that the Lawsuit messages actually appear as follow-up tricks of Financial scams. We term this fraudulent model as *Semantic Progressiveness* and would describe it in detail later.

We also measure the lifetime of a campaign by the number of days to send spearphishing messages in this work. The average lifetime was 3 days for all campaigns and 31 days for the top 100. Interestingly, the lifespan also varies among different types of spearphishing businesses. Lawsuit Scam and Gambling Phishing campaigns have the longest average lifespan, at 7 days and 6 days respectively. In contrast, Financial Scam campaigns survived the shortest, with an average lifetime of 2 days. From the perspective of evasion, the above phenomenon could be reasonable: Contacts left in Financial Scam messages may be directly involved in monetary transaction operations with victims, resulting in them being at high-risk and subject to more frequent replacement. Therefore, as we distinguish campaigns by embedded contacts, these campaigns would present shorter lifetime.

Furthermore, our exploration reveals several interesting strategies of spearphishing attacks. To evade detection, they perform test-sending on controlled devices and use multi-semantic text to hide their true purpose. To attract victims, they construct phishing content with global trending affairs and devise a chain of scams to deceive victims. These strategies are described in detail as follows.

Strategy of spearphishing: test-sending. We observed an interesting phenomenon, that one device (with a unique IMSI-IMEI) continuously received 219 Lawsuit Scam spearphishing spam messages from the same campaign (see one example of Lawsuit Scam in Table 3). The messages were templated, with only the Names in the text changing (211 unique Names observed). Besides this

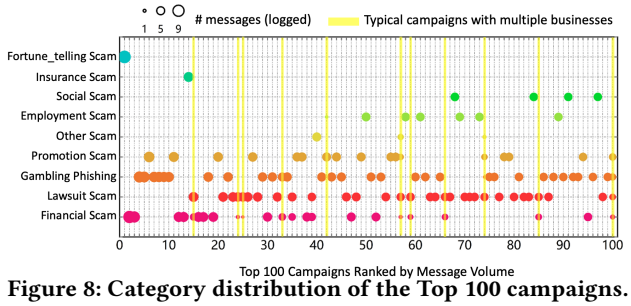


Figure 8: Category distribution of the Top 100 campaigns.

"abnormal" victim, the company also sent spearphishing messages to several other "normal" victims (each received only one message) using similarly templated content. After discussions with telecom-fraud experts, we speculate that this is an interesting strategy, test-sending operation, of spearphishing attackers. As both the Internet Service Provider (ISP) and the client-side application may implement fraud blocking features, spearphishing messages sent by attackers may not reach the user's inbox at all. Therefore, attackers would check the "passing rate" by sending test messages to a device under their control (as the "abnormal victim" we observed) and then select the "passable messages" to spoof "normal victims" on a large-scale, which is denoted as the test-sending strategy.

To understand the prevalence of this strategy in spearphishing ecosystem, we systematically detect it across the entire dataset. Campaigns with the following two characteristics would be screened out: 1) At least one victim received multiple templated spearphishing messages with different names as a testing device. 2) In addition to the testing device, messages from similar templates were also sent to multiple normal victims (each received messages with a unique name). In this way, 4 campaigns employing the "test-sending" strategy are identified, which totally sent 539 test messages and 9,062 formal spearphishing messages during the 3-month data-collection period and affected 2,275 victims.

The above findings could provide new insights for defense, i.e., in addition to detecting spearphishing messages based on text content in real-time, we could also consider identifying test-sending operations of attackers (feasible for ISPs as they own the sender-receiver relationship logs), which could be helpful to block spearphishing attacks early in the test-sending phase.

Strategy of spearphishing: progressive deception. In addition to the devices for test-sending mentioned above, we also find that 12,752 (52.11%) victims received multiple spearphishing messages from one single campaign, with 192 of whom exhibiting another interesting phenomenon: the received messages were from multiple spearphishing categories with the content showing a semantic progression. For example, 122 victims first received pieces of Financial Scam messages, in which they were recommended to subscribe to loan service. Sometime later, we observed that all these victims received Lawsuit Scam messages from the same campaign, claiming the loan had expired and tricking them to contact with the attackers, or they would be prosecuted for the debt. We speculate these victims may have indeed used the loan services in the first round of the scam, and they may have continued to be trapped in the second round for the fear of prosecution. We term this phenomenon as progressive deception, in which the attackers do not simply commit

[Guangfa Bank] Dear Alice: Bob remitted 2.88 RMB to your account with the message: *Tiancai Gambling service in tc***.com will send bonus to new users*.

Figure 9: Example of Multi-semantic Evasion: the attacker hides a Gambling Phishing message (in red) in the text with financial semantics.

the spearphishing multiple times, but craft a chain of scams to perpetuate the attack in a long-term and systematic manner. Although it is more costly than one-time scams, i.e., the strategy requires continuous tracking the individual status of victims (over 40% of the victims were tracked for more than 10 days, with the longest being 85 days from our data), this carefully constructed scam is also more deceptive to the victims and would be taken by the attackers for the potentially high profits.

Strategy of spearphishing: multi-semantic evasion. During manual inspections of Financial Scam messages, we find several "hidden false positives", one example of which is shown in Figure 9. In fact, it was used to advertise gambling, with the coded domain pointing to a Gambling Phishing website. However, as the gambling-related text only appears as a comment of a bank transfer, it was able to be hidden in the semantics of normal financial services and is indeed identified "incorrectly" by our multi-classifier as Financial Scam (rather than a Gamling Phishing). We consider it to be a new scam strategy, which we termed as multi-semantic evasion, where the attacker embeds text with suspicious semantics into a relatively normal SMS to hide its true purpose.

We also tried to detect this behavior across the entire dataset. In order to find messages with multiple semantics, we split a message at the middle into two parts, and then identify their semantic categories separately. One message could be marked as suspicious if the categories of its two parts did not match, and then a manual check would be processed to ensure the mix of semantics is for evasion. In the end, we identified 1,197 spear spam messages of multi-semantic evasion from 416 campaigns, covering 839 victims, and 706 of them have successfully hidden their true purposes in the content identification of our multi-classifier.

Strategy of Spearphishing: global affair integration. It is worth noting that, a highly influential public event, the COVID-19 virus, emerged just during our data collection period (Dec. 2019 to Mar. 2020). As public reports have identified cyber-crimes associated with COVID-19 [63], in the field of spearphishing SMS, we also discovered **276** COVID-19 related cases. Attackers exploited global concerns and fears about the virus to lure users. For example, they threaten victims that their invested funds would be withheld due to the effects of the virus and provide a cellphone for victims to contact further. In another case, attackers used the masks, a scarce medical item at that time, as bait to entice users to sign up for gambling websites with an example of Gambling Phishing in Table 3.

5.4 Real-world Impact of SMS Spearphishing

In this section, we evaluate the impact of spearphishing SMS attacks from the following three perspectives.

Victim Coverage. The detected spearphishing spam messages were related with 24,472 IMSI (unique identifier of SIM Card) and 24,346 IMEI (unique identifier of Mobile Equipment) (both have been anonymized by hashes). In other words, around 24k victims were endangered by spearphishing spam SMS from Dec. 2019 to Mar. 2020, covering all the provinces in China.

"Looks Safe" Rate. As mentioned in Section 3, users could recover the detected messages to Normal Inbox, when those messages are actually useful to them. Our industrial partner provided tags for these messages that have been recovered as "looks safe". We found the "looks safe" rate of spearphishing is 0.04%, which is four times over common spam ones (0.01%). It indicates that the messages embedded with the user's personal information can attract more attention, which also means the success rate of spearphishing attacks is higher than common spam to trap users.

Follow-up Domain Visits. As shown in Table 2, more than 60% detected messages embed URL/Domain. The visits of these domains during the data collection period could help to value the actual impact of spearphishing attacks. Here we utilize the Passive DNS database of 360 Netlab[5], which has better coverage in China than other PDNS data sources as DNSDB[18]. From the 1,473 domains in spearphishing SMS we detected, 80% of them received more than 100 queries and 11% had been visited more than 5,000 times during the data collection period. Factoring the impact of spearphishing, we found 87.62% of requests occurred after the spearphishing messages were sent. Moreover, 1,392 (94.50%) domains are marked as malicious by at least one threat intelligence [1, 57], of which 105 are phishing-related. In particular, one phishing domain appeared in 432 messages embedding victims' License Plate Numbers. In this case, the attackers disguised as vehicle authorities to lure victims into clicking on the phishing domain, which actually redirected to a gambling scam website. PDNS logs revealed that, this scam website received 5,136 requests after spearphishing messages were sent, with an increase of 37.2% in average daily requests. More seriously, we find that 28 domains are malware-related, and the spearphishing messages increased their daily request volume by an average of 27.03%. Despite the inherent limitations of PDNS data, we believe the above findings are sufficient to confirm spearphishing SMS attacks have made a considerable impact in practice.

To sum up, we observed 24k victims of spearphishing SMS attacks in 3 months, covering every province inside China. Spearphishing messages are more likely to be recognized as "normal messages" from the user's perspective. Therefore, we suggest that spearphishing attacks in SMS do cause serious harm in the real world.

5.5 Personal Info in Spearphishing SMS

In this section, we discuss the leaked personal information leveraged in spearphishing SMS, including how attackers use them to build customized scam content and the possible sources of leakage.

Through manual inspections, we find Names are usually placed at the beginning of the SMS with a salutation or greeting to attract the victims' attention. While after getting the Flight Info (883 messages), attackers would pose as staff members of airlines, claiming the flight has been canceled or delayed, and leave one private cellphone in the message for subsequent scams. In this study, we find several major Chinese airlines, including China Eastern Airlines, Shanghai Airlines and Air China, have been affected. License Plates (571 messages) are commonly used to impersonate car insurance companies, asking victims to check their insurance status via embedded URLs that actually pointed to gambling phishing websites

or distributed malicious APKs. Besides, IDs (17 messages) are also exploited for financial fraud.

How exactly the victims' information gets leaked is a matter of great concern. In the field of telecom scams, previous work [6] has identified social networks, malware, and public forums as the possible sources of information leakage. In this work, it is difficult to track and precisely locate the source of detected PIIs on a large scale (especially when researchers do not have direct access to the actual content of leaked information). However, we did find some possible sources through manual case studies.

First, we find 10,598 Employment messages, of which 9,027 were impersonation popular online job websites in China [56]. In these messages, attackers falsely claimed that the victim's job request was approved and lured the victim into further scams through follow-up contacts. A 2019 report announced that third parties may collect PIIs [23] by capturing resumes uploaded by job seekers on such websites. We also perform manual checks on the recruitment website, and find that it was indeed possible to access the personal information of job applicants through their CVs.

Second, we find 274 Insurance Scam messages impersonating *China Post*, where the attackers claim that a certain insurance product purchased by victims has expired and lure the victims to renew the policy by clicking on one embedded link that actually redirects to a Gambling Phishing website. Interestingly, all the victims of this case were concentrated in the provinces of Guangdong (41.97%) and Guangxi (30.66%), and the spoofed product was also served for users in that regions. Therefore, we speculate that the personal information of victims in this incident may have originated from internal leaks within several local organizations.

Limitation of Verification. As described in Section 3.2, all detected personal information of victims has been hashed, with only the type and corresponding hash value given to researchers. It maximizes the protection of the user privacy, while also makes it hard to perform corresponding validations of PIIs. In other words, we are unable to validate whether the name embedded in spearphishing messages matches with the real name of the victim, not to mention further tracing the leakage source of PIIs. It is considered as one major limitation of evaluation in this work. However, even in the relatively mature research area of spearphishing emails, few existing works could validate whether the information in fraudulent email content exactly matches the actual personal information of the victim [25, 26, 55]. Besides, the authentication of Name could only affect the success rate of attacks, without changing the fact that the messages we detected are customized, fraudulent, and sent to specific victims, i.e., compliant with the definition of spearphishing. Moreover, we focus on detecting the occurrence of spearphishing SMS attacks and understanding the behind strategies, rather than studying their effectiveness (success rate). The validation of the accuracy of victim's PIIs could be explored in the future through the proactive deployment of honeypots as previous work [6].

6 DISCUSSION

Comparison with Spearphishing via other channels. As one of the social engineering attacks where the attackers pretend as trusted senders and send customized phishing content, spearphishing via SMS has unique characteristics compared with other channels. First, cellphones are more commonly used than emails and social software, increasing the risk of relevant information being leaked to attackers. Second, while SMS is not richly formatted, limiting the extent to which attackers could customize the phishing content, it also raises difficulties to detection. Previous detection of spearphishing on other channels based on rich field information, such as headers and forward relationships, could not be directly applied to SMS. Thus, we proposed a novel detecting system in this work. Our results corroborate similarities between spearphishing SMS and attacks via other channels, e.g., both require the persistent interaction with potential victims [25], and also discover unique features of spearphishing SMS such as "test-sending" and "multi-semantic" strategies.

Comparison with SMS Spam. Previous works on spam have explored this ecosystem primarily in terms of different distribution channels, such as spam from SMS Gateways [48, 49] and spam from Fake Base Stations [73]. However, even if from the same channel, the practical security risks of different SMS content are quite varied. For example, while ordinary promotional SMS can be at best "annoying" users, phishing messages, especially the high-risk spearphishing messages, would expose users to serious information leakage or property damage. This work investigates the spearphishing attack, which probably is the most "high-risk" part of the entire spam SMS ecosystem. We believe our findings would provide assistance in addressing the core issues in the field of SMS spamming.

Recommendation. Bootstrapping from our measurement findings, we provide several recommendations to mitigate this security threat. First, our proposed spearphishing detection system would enable mobile security applications to improve their detection capabilities. They could also provide eye-catching risk alerts of spearphishing SMS by revising the UI design of notification. Second, as SMS gateways are being abused as the major channel for sending spear messages, ISPs could consider cooperating with security vendors to enhance the audits of content submitted to their SMS platforms. Furthermore, the evasion strategies of spearphishing observed in this study can provide new insights for ISPs for detection, such as monitoring and identifying test-sending patterns and blocking large-scale spear attacks at the initial phase. Besides, although we only analyzed data collected within China, our methods, such as the use of sentence structure to detect spear spam, are also worth replicating in spear detection in other regions.

7 RELATED WORK

Spearphishing Attack. Previous works mainly focused on the detection of spearphishing attacks, especially for spearphishing emails. Attackers usually utilize account spoofing for spearphishing, which could be classified into external attacks, where the attackers need to imitate some well-known accounts (or one known to the victim) to imbue their profile with a sense of trust or authority [26], and internal attacks, where attackers are more insidious as they send malicious content to victims by getting control of compromised accounts [25, 55]. Proposed detection methods of spearphishing emails rely on features of embedded URLs [9, 20], content of webpages [64, 65, 72], and linguistic features of the email headers and body content [2, 11, 17, 26, 35, 69]. In recent years, telecommunications have also been abused as the main channels for scams [38], with the emerging attacking techniques like caller ID spoofing [59].

Our work gives the first large-scale detection and characterization of spearphishing attacks via SMS channel.

Spam SMS. Previous works have explored spam SMS sent through a variety channels, including fake base stations [33, 47, 73] or SMS Gateways [48, 49]. Existing approaches for detecting spam SMS are mainly include template-based clustering [4, 13, 22], topic analysis [36] and clustering based on the sending behaviors of suspicious accounts [32]. However, there are no works that discussed the emerging new threat, *SMS spearphishing attack*, which utilizes victims' personal information to construct deceptive content.

8 CONCLUSION

In this work, we first explored a new threat, SMS spearphishing attack, through the three-month real-world dataset (31.97 million) in China. We designed and implemented a novel detection system based on the three key observations obtained from an empirical study, and detected a total of 90,801 SMS spearphishing messages on the whole dataset with 96.16% precision. Measurement of those detected messages revealed multi-faceted characteristics of SMS spearphishing attacks, like business categories, temporal characteristics and spatial characteristics, and the infrastructures of attackers. Besides, by grouping the messages into 11,475 campaigns, we firstly found several interesting strategies of attackers and provided a comparative analysis with other types of spam. Our findings would assist the security community in understanding and mitigating SMS spearphishing attacks.

ACKNOWLEDGMENTS

We thank our shepherd Gianluca Stringhini and all the anonymous reviewers for their valuable comments to improve this paper. This work was supported in part by the National Natural Science Foundation of China under Grant U1836213, the National Key Re-search and Development Program of China under Grant 2018YFB2101501, Ministry of Industry and Information Technology of China under Grant TC200H02Y and TC200H02X. Baojun Liu is partially supported by the NSFC 62102218 and Shuimu Tsinghua Scholar Program. Zhou Li is partially supported by gift from Microsoft and Cisco.

REFERENCES

[1] Qihoo 360. Accessed May, 2020. Threat intelligence platform in Qihoo 360. https://ti.360.cn.

[2] Shivam Aggarwal, Vishal Kumar, and SD Sudarsan. 2014. Identification and detection of phishing emails using natural language processing techniques. In *Proceedings of the 7th International Conference on Security of Information and Networks*. 217–222.

[3] Mark Alllman and Vern Paxson. 2007. Issues and etiquette concerning use of shared measurement data. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*. 135–140.

[4] Tiago A Almeida, José María G Hidalgo, and Akebo Yamakami. 2011. Contributions to the study of SMS spam filtering: new collection and results. In *Proceedings of the 11th ACM symposium on Document engineering*. 259–262.

[5] Netlab at Qihoo 360. Accessed May, 2020. Passive DNS System. https://passivedns.cn/.

[6] Marco Balduzzi, Payas Gupta, Lion Gu, Debin Gao, and Mustaque Ahamad. 2016. Mobipot: Understanding mobile telephony threats with honeycards. In *Proceedings of the 11th ACM on Asia Conference on Computer and Communications Security*. 723–734.

[7] Becky Bracken. 2021. LinkedIn Spear-Phishing Campaign Targets Job Hunters. https://threatpost.com/linkedin-spear-phishing-job-hunters/165240/.

[8] China CCTV. 2019. Ranking of permanent residents in 31 provinces in China. https://news.cctv.com/2019/06/16/ARTIESM4vkQakTiZC8YxeA24190616.shtml.

[9] Juan Chen and Chuanxiong Guo. 2006. Online detection and prevention of phishing attacks. In *2006 First International Conference on Communications and Networking in China*. IEEE, 1–7.

[10] Science China. 2017. 360 Mobile safe's coverage rate is far ahead, ranking first in security software in China. http://science.china.com.cn/2017-08/17/content_39086982.htm.

[11] Asaf Cidon, Lior Gavish, Itay Bleier, Nadia Korshun, Marco Schweighauser, and Alexey Tsitkin. 2019. High precision detection of business email compromise. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1291–1307.

[12] Xiamen Wanderlust Technology Co. 2014. National public service hotlines. https://www.ip138.com/tel.htm.

[13] Gordon V Cormack, José María Gómez Hidalgo, and Enrique Puertas Sánz. 2007. Spam filtering for short messages. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. 313–320.

[14] Baris Coskun and Paul Giura. 2012. Mitigating sms spam by online detection of repetitive near-duplicate messages. In *2012 IEEE International Conference on Communications (ICC)*. IEEE, 999–1004.

[15] Katie DeMatteis. 2019. What's So Dangerous About Spear Phishing? https://www.carbonblack.com/blog/whats-so-dangerous-about-spear-phishing/.

[16] David Dittrich, Erin Kenneally, et al. 2012. *The Menlo Report: Ethical principles guiding information and communication technology research*. Technical Report. US Department of Homeland Security.

[17] Sevtap Duman, Kubra Kalkan-Cakmakci, Manuel Egele, William Robertson, and Engin Kirda. 2016. Emailprofiler: Spearphishing filtering with header and stylometric features of emails. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 1. IEEE, 408–416.

[18] FarSight Security. Accessed May, 2020. DNSDB data. https://dnsdb.io/zh-cn/.

[19] Food and Agriculture Organization of the United Nations. [n.d.]. Country code and area code. http://www.fao.org/countryprofiles/iso3list/zh/. Accessed June, 2020.

[20] Sujata Garera, Niels Provos, Monica Chew, and Aviel D Rubin. 2007. A framework for detection and measurement of phishing attacks. In *Proceedings of the 2007 ACM workshop on Recurring malcode*. 1–8.

[21] Alexander Genkin, David D Lewis, and David Madigan. 2007. Large-scale Bayesian logistic regression for text categorization. *technometrics* 49, 3 (2007), 291–304.

[22] José María Gómez Hidalgo, Guillermo Cajigas Bringas, Enrique Puertas Sánz, and Francisco Carrero García. 2006. Content based SMS spam filtering. In *Proceedings of the 2006 ACM symposium on Document engineering*. 107–114.

[23] Hacken. 2019. NO MORE PRIVACY: 202 MILLION PRIVATE RESUMES EXPOSED. https://hacken.io/research/industry-news-and-insights/no-more-privacy-202-million-private-resumes-exposed/.

[24] Han He. 2020. *HanLP: Han Language Processing*. https://github.com/hankcs/HanLP

[25] Grant Ho, Asaf Cidon, Lior Gavish, Marco Schweighauser, Vern Paxson, Stefan Savage, Geoffrey M Voelker, and David Wagner. 2019. Detecting and characterizing lateral phishing at scale. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1273–1290.

[26] Grant Ho, Aashish Sharma, Mobin Javed, Vern Paxson, and David Wagner. 2017. Detecting credential spearphishing in enterprise settings. In *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 469–485.

[27] Longtao Huang, Ting Ma, Junyu Lin, Jizhong Han, and Songlin Hu. 2019. A Multimodal Text Matching Model for Obfuscated Language Identification in Adversarial Communication?. In *The World Wide Web Conference*. 2844–2850.

[28] M Ikonomakis, Sotiris Kotsiantis, and V Tampakas. 2005. Text classification using machine learning techniques. *WSEAS transactions on computers* 4, 8 (2005), 966–974.

[29] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*. 1188–1196.

[30] Jinfeng Li, Tianyu Du, Shouling Ji, Rong Zhang, Quan Lu, Min Yang, and Ting Wang. 2020. TextShield: Robust Text Classification Based on Multimodal Embedding and Neural Machine Translation. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*.

[31] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271* (2018).

[32] Rui Li, Yongzheng Zhang, Yupeng Tuo, and Peng Chang. 2018. A novel method for detecting telecom fraud user. In *2018 3rd International Conference on Information Systems Engineering (ICISE)*. IEEE, 46–50.

[33] Zhenhua Li, Weiwei Wang, Christo Wilson, Jian Chen, Chen Qian, Taeho Jung, Lan Zhang, Kebin Liu, Xiangyang Li, and Yunhao Liu. 2017. FBS-Radar: Uncovering Fake Base Stations at Scale in the Wild.. In *NDSS*.

[34] Zihan Liu, Yan Xu, Tiezheng Yu, Wenliang Dai, Ziwei Ji, Samuel Cahyawijaya, Andrea Madotto, and Pascale Fung. 2020. CrossNER: Evaluating Cross-Domain Named Entity Recognition. *arXiv preprint arXiv:2012.04373* (2020).

[35] Christian Ludl, Sean McAllister, Engin Kirda, and Christopher Kruegel. 2007. On the effectiveness of techniques to detect phishing sites. In *International Conference on Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 20–39.

[36] Jialin Ma, Yongjun Zhang, Jinling Liu, Kun Yu, and XuAn Wang. 2016. Intelligent SMS spam filtering using topic model. In *2016 International Conference on*

Intelligent Networking and Collaborative Systems (INCoS). IEEE, 380–383.
[37] Huanguo Message. 2021. The Huanguo Messages for 106 SMS Platform. http://www.106.cn/product.
[38] Najmeh Miramirkhani, Oleksii Starov, and Nick Nikiforakis. 2016. Dial one for scam: A large-scale analysis of technical support scams. *arXiv preprint arXiv:1607.06891* (2016).
[39] China Mobile. 2021. Mobile cloud services from China Mobile. https://saas.ecloud.10086.cn/Store/TSDetail/1524.
[40] Ilona Murynets and Roger Piqueras Jover. 2012. How an SMS-Based malware infection will get throttled by the wireless link. In *2012 IEEE International Conference on Communications (ICC)*. IEEE, 960–965.
[41] China News. 2016. 360 Mobile safe has the largest number of users in China. https://china.huanqiu.com/article/9CaKrnJZhMn.
[42] China News. 2021. GDP of 31 provinces in 2021. http://www.xinhuanet.com/fortune/2021-04/28/c_1127386550.htm.
[43] Economic Reference News. 2016. New telecom scam scheme with Personal information dumping has become a black industry chain. http://finance.people.com.cn/n1/2016/0909/c1004-28703097.html.
[44] Xinhua News. 2013. Top Ten Types of Spam Messages Real Estate Advertising Becomes the "King of Spam Messages". http://media.people.com.cn/n/2013/1016/c40733-23222153.html.
[45] Xinhua News. 2018. What's behind the of 106 nuisance SMS? http://www.xinhuanet.com/2018-12/16/c_1123860405.htm.
[46] Federal Bureau of Investigation (FBI). 2018. Business E-mail Compromise The 12 Billion Dollar Scam. https://www.ic3.gov/Media/Y2018/PSA180712.
[47] Sohu Media Platform. 2016. Demystifying the Industrial Chain of Fake Base Stations. http://m.sohu.com/n/444726367/.
[48] Bradley Reaves, Logan Blue, Dave Tian, Patrick Traynor, and Kevin RB Butler. 2016. Detecting SMS spam in the age of legitimate bulk messaging. In *Proceedings of the 9th ACM Conference on Security & Privacy in Wireless and Mobile Networks*. ACM, 165–170.
[49] Bradley Reaves, Nolen Scaife, Dave Tian, Logan Blue, Patrick Traynor, and Kevin RB Butler. 2016. Sending out an SMS: Characterizing the Security of the SMS Ecosystem with Public Gateways. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 339–356.
[50] 360 Mobile Safe. 2020. Report on China's mobile phone security in the first half of 2020.
[51] 360 Mobile Safe. Accessed May, 2020. The privacy policy of 360 Mobile Safe. http://shouji.360.cn/about/privacy/index_2.0.html.
[52] John Seymour and Philip Tully. 2018. Generative models for spear phishing posts on social media. *arXiv preprint arXiv:1802.05196* (2018).
[53] Ge Song, Yunming Ye, Xiaolin Du, Xiaohui Huang, and Shifu Bie. 2014. Short text classification: A survey. *Journal of multimedia* 9, 5 (2014), 635.
[54] China Telecom. 2021. SMS Group Sending Platform-Three Network Jiexin 106 SMS Platform. http://www.106vip.net/783.html.
[55] Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. 2013. Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse. In *22nd {USENIX} Security Symposium ({USENIX} Security 13)*. 195–210.
[56] 58 Tongcheng. Access June, 2020. 58 Tongcheng: Providing biggest Free information classifieds service in China. https://58.com/.
[57] Virus Total. Access May, 2020. Virus Total. https://www.virustotal.com/gui/home/search.
[58] Bruno Trstenjak, Sasa Mikac, and Dzenana Donko. 2014. KNN with TF-IDF based framework for text categorization. *Procedia Engineering* 69 (2014), 1356–1364.
[59] Huahong Tu, Adam Doupé, Ziming Zhao, and Gail-Joon Ahn. 2019. Users really do answer telephone scams. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1327–1340.
[60] International Telecommunications Union. 2014. Country code and area code. https://www.itu.int/dms_pub/itu-t/opb/sp/T-SP-M.1400-2014-PDF-C.pdf.
[61] Mike Vizard. 2019. Mueller Report details how long national nightmare started with simple spearphishing campaign. https://blog.barracuda.com/2019/04/26/mueller-report-details-how-long-national-nightmare-started-with-simple-spearphishing-campaign/.
[62] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. 2017. Combining Knowledge with Deep Convolutional Neural Networks for Short Text Classification.. In *IJCAI*, Vol. 350.
[63] Kevin Watkins. 2020. SMS Phishing Campaigns Take Advantage of Coronavirus Pandemic. https://symantec-enterprise-blogs.security.com/blogs/threat-intelligence/sms-phishing-coronavirus.
[64] Colin Whittaker, Brian Ryner, and Marria Nazif. 2010. Large-scale automatic classification of phishing pages. (2010).
[65] Guang Xiang, Jason Hong, Carolyn P Rose, and Lorrie Cranor. 2011. Cantina+ a feature-rich machine learning framework for detecting phishing web sites. *ACM Transactions on Information and System Security (TISSEC)* 14, 2 (2011), 1–28.
[66] Yanqian Xu. 2016. Civil Aviation Authority requires airlines' websites to warn against SMS scams, experts call for higher costs of breaking the law. https://www.thepaper.cn/newsDetail_forward_1465286.
[67] Hao Yang, Kun Du, Yubao Zhang, Shuang Hao, Zhou Li, Mingxuan Liu, Haining Wang, Haixin Duan, Yazhou Shi, Xiaodong Su, et al. 2019. Casino royale: a deep exploration of illegal online gambling. In *Proceedings of the 35th Annual Computer Security Applications Conference*. 500–513.
[68] Hao Yang, Xiulin Ma, Kun Du, Zhou Li, Haixin Duan, Xiaodong Su, Guang Liu, Zhifeng Geng, and Jianping Wu. 2017. How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy. In *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 751–769.
[69] Weider D Yu, Shruti Nargundkar, and Nagapriya Tiruthani. 2009. Phishcatch-a phishing detection tool. In *Proceedings of the 2009 33rd Annual IEEE International Computer Software and Applications Conference-Volume 02*. 451–456.
[70] Natalya Zablotskaya. 2008. Fraudulent spam. https://securelist.com/fraudulent-spam/36218/.
[71] Shengnan Zhang, Yan Hu, and Guangrong Bian. 2017. Research on string similarity algorithm based on Levenshtein Distance. In *2017 IEEE 2nd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*. IEEE, 2247–2251.
[72] Yue Zhang, Jason I Hong, and Lorrie F Cranor. 2007. Cantina: a content-based approach to detecting phishing web sites. In *Proceedings of the 16th international conference on World Wide Web*. 639–648.
[73] Yiming Zhang, Baojun Liu, Chaoyi Lu, Zhou Li, Haixin Duan, Shuang Hao, Mingxuan Liu, Ying Liu, Dong Wang, and Qiang Li. 2020. Lies in the Air: Characterizing Fake-base-station Spam Ecosystem in China. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*.
[74] Yue Zhang and Jie Yang. 2018. Chinese NER using lattice LSTM. *arXiv preprint arXiv:1805.02023* (2018).
[75] Zihan Zhang, Mingxuan Liu, Chao Zhang, Yiming Zhang, Zhou Li, Qi Li, Haixin Duan, and Donghong Sun. 2020. Argot: Generating Adversarial Readable Chinese Texts. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2533–2539. https://doi.org/10.24963/ijcai.2020/351 Main track.

A EXAMPLE OF REGULAR EXPRESSION

Table 4: The example of regular expression of each PIIs.

Type	Pattern Description	Regular Expression
Flight	2 letters of Airline Codes and 4 digits or letters.	[a-zA-Z]{2}[0-9]{3}[a-zA-Z0-9]
License Plate	1 Chinese character, 1 letter and 5 digits or letters.	[Provice_code_list]{1}[A-Z]{1}[A-HJ-NP-Z0-9]{5}
ID Card	Version 1: 6-digit address code, 6-digit date of birth code and 3-digit sequential code (totally 15 digits).	[0-9]{6}[0-9]{6}[0-9]{3}
	Version 2: 6-digit address code, 8-digit date of birth code, 3-digit sequential code and 1 check digit or letter (totally 18 digits).	[0-9]{6}[0-9]{8}[0-9]{3}[0-9XY]{1}
Bank Card	16-20 digits	[0-9]{19}

B CLASSIFICATION PERFORMANCE

Table 5: The performance of each classification model

Classifier	Word2Vec			TF-IDF		
	pr	rc	f1	pr	rc	f1
LR	93.46%	93.47%	93.42%	93.17%	93.18%	93.04%
DT	86.39%	86.40%	86.26%	93.16%	93.15%	93.04%
Bernoulli NB	75.76%	75.77%	75.63%	92.98%	92.99%	92.96%
Gauss NB	74.95%	74.95%	75.47%	90.96%	90.99%	91.01%
SVM	61.54%	61.53%	61.33%	61.68%	61.68%	60.70%

[1] Classifier: LR: Logistic Regression, DT: Decision Tree, Bernoulli NB: Naive Bayes, Gauss NB: Naive Bayes, SVM: Support Vector Machine.
[2] Evaluation indicators: pr: precision, rc: recall, f1: f1 score.