Understanding and Detecting Private Interactions in Underground Forums

Zhibo Sun[1], Carlos E. Rubio-Medrano[1], Ziming Zhao[2], Tiffany Bao[1]
Adam Doupé[1], Gail-Joon Ahn[1,3]
[1] Arizona State University
{zhibo.sun,crubiome,tbao,doupe,gahn}@asu.edu
[2] Rochester Institute of Technology
zhao@mail.rit.edu

ABSTRACT

The studies on underground forums and marketplaces have significantly advanced our understandings of cybercrime workflows and underground economies. Researchers of underground economies have conducted comprehensive studies on public interactions. However, little research focuses on private interactions. The lack of the investigation on private interactions may cause misunderstandings on underground economies, as users in underground forums and marketplaces tend to share the minimal amount of information in public interactions and resort to private messages for follow-up conversations.

In this paper, we propose methods to investigate the underground private interactions and we analyze a recently leaked dataset from Nulled.io. We present analyses on the contents and purposes of private messages. In addition, we design machine learning-based models that only use the publicly available information to detect if two underground users privately communicate with each other. Finally, we perform adversarial analysis to evaluate the robustness of the detector to different types of attacks.

CCS CONCEPTS

• **Security and privacy → Social network security and privacy**; • **Information systems** → *Deep web*.

KEYWORDS

Underground forums, private interaction analysis, private interaction detection

ACM Reference Format:
Zhibo Sun[1], Carlos E. Rubio-Medrano[1], Ziming Zhao[2], Tiffany Bao[1], Adam Doupé[1], Gail-Joon Ahn[1,3]. 2019. Understanding and Detecting Private Interactions in Underground Forums. In *Ninth ACM Conference on Data and Application Security and Privacy (CODASPY '19), March 25–27, 2019, Richardson, TX, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3292006.3300036

1 INTRODUCTION

Underground forums and marketplaces have been the rendezvous sites for cybercriminals of all kinds to exchange information and sell illegal products and services. Given the important roles these sites play in the cybercrime ecosystem, a considerable amount of research effort has been invested in studying the organizational structures of their users [2, 46, 49], their social dynamics, such as how users gain and lose trust [1, 35], the goods and services sold [13, 17, 44, 47], and how such sites assist specific forms of cybercrimes [21, 43]. These studies have significantly advanced our understandings of the underground economies and have guided law enforcement agencies and affected businesses in how to respond cybercrimes [44].

Nevertheless, the research community has only been looking at the tip of the iceberg of underground forums and marketplaces in that *we have only investigated posts and threads that are publicly available to all registered users or even anyone on Internet.* However, due to the nature of underground forums and marketplaces, their users tend to share a very small amount of information in public interactions and resort to private messages for follow-up conversations. Hence, analyzing private messages can disclose a wealth of information, such as the illicit financial flow, narrowing down the suspects who commit the crime, etc., while our community is not even aware of what sort of information we could glean from private messages.

In this paper, we investigate private messages in underground forums by analyzing the leaked dataset of Nulled.io. Nulled.io is a popular underground forum where users discuss hacking, exploits, monetization methods, etc. The released forum dataset includes public and private messages from Jan 14, 2015 to May 6, 2016, which makes it an excellent sample for comparing the similarities and differences between public and private interactions.

To study the private messages of Nulled.io, we develop a semi-automatic approach to categorize private messages into content and purpose categories, to compare how private messages are different from public posts. We present the artifacts discovered via our analysis of private messages, such as the payment methods, contact information, etc. Our analyses show that the content and purpose distributions in private messages *are different from* their public counterparts and that there is much more sensitive information, such as the users' contact information and Bitcoin addresses in private messages than in public posts. In the meantime, private exchanged information is not always undisclosed in public.

[3] Dr. Gail-Joon Ahn is also affiliated with Samsung Research.

CODASPY '19, March 25–27, 2019, Richardson, TX, USA

ACM ISBN 978-1-4503-6099-9/19/03…$15.00
https://doi.org/10.1145/3292006.3300036

In addition, we analyze who are more likely to be contacted in private messages by studying public interactions of `recipients`, users who receive the initial message in a private thread, to understand how they behave differently from other users. We compare public and private interactions of `recipients` to study the relationships between their public and private activities. Also, we analyze different types of public posting methods, such as creating a post and replying to a post, and study which type of posting method is more likely to attract private interactions.

We also design various machine learning-based approaches to *detect private interactions* in underground forums based solely on public information. In essence, our approach uses publicly available information to uncover hidden connections between users. The evaluation results indicate that our approach effectively detects private interactions with 94% accuracy. Additionally, we perform adversarial analysis to evaluate the robustness of the detector to different types of attacks. Our analyses indicate that our detection has a considerable effective robustness to multiple types of evasion attacks. Also, our evaluations indicate that poisoning attacks performed by the administrators cannot prevent the private interactions from the detector effectively.

The contributions of this paper are summarized as follows:

- We analyze private messages in the underground forum in terms of different types of discussed content and purpose to compare how private messages are different from public posts. Also, we manually analyze the most popular artifacts in private messages.
- We study the public activities of private message recipients to understand how they behave differently from general users. Also, we analyze the relationship between public and private interactions of private message recipients.
- By considering the characteristics of underground forums and leveraging findings from previous studies and carefully selected features, we can effectively detect private interactions from public information.
- We present an adversarial analysis against our detector to understand the robustness of our detection algorithm.

2 MOTIVATING EXAMPLE AND DATA OVERVIEW

There are many studies on analyzing user interactions in public social networks [7, 8, 26, 31], most of which focus on public interactions that can be accessed by anyone, such as likes, reposts, shares, replies or tagging pictures. Similar public interactions are also utilized to analyze the social dynamics of underground forum users [33, 46, 49]. However, private messages on underground forums have been largely overlooked. In this section, we present a motivating example that sheds light on what may be discovered in private messages. We also discuss the dataset that is used in this paper and how we preprocess the raw data.

2.1 Motivating Example

As a motivating example, as shown in Figure 1, a vendor of Nulled.io tries to sell cracked accounts in a public post. However, the public post does not include any detailed information that is needed to finish a transaction, such as the payment or delivery methods. In contrast, corresponding private messages from the same vendor is shown in Figure 2, and these message comprise much more detailed information. To anonymize users' identities, we use "Buyer" and "Vendor" to represent their user IDs in the forum. We also use "BTCADDRESS," "USERNAME," and "PASSWORD" to represent the vendor's Bitcoin address and the cracked account login credentials.

```
Vendor: Selling HideMyAss VPN premium accounts. Type of accounts:
      monthly membership with auto renew. Price is $5 BTC. Do not
      change the password/email of the account. I am not responsible
      if you get banned for breaking the TOS of HMA, make sure to
      read it first. All sales are final.
```

Figure 1: A public post in which a vendor is selling cracked accounts

```
Buyer: I want to buy one hide my ass account with bitcoin can I?
      Thanks
Vendor: Sure, price is $5 BTC. Ready to send BTC over? BTW i can
      give you 3 accounts for $12 BTC, that is a current deal. :)
Buyer: No I want only one thanks. Yes, I'm ready please sen me your
      btc address.
Vendor: Sure, send btc to: <BTCADDRESS>
Buyer: Sent.
Vendor: <USERNAME>:<PASSWORD>. Enjoy :)
Vendor: BTW can you leave a vouch in my sales thread? Thanks in
      advance. :)
Buyer: Thank you too. Feedback sent :)
```

Figure 2: Private messages exchanged between the same vendor in Figure 1 and a buyer. The bitcoin address, username, and password have been redacted.

The private messages in Figure 2 show that a user contacts the vendor to buy a cracked account of *HideMyAss*, which is a VPN service provider and its original service price is $11.99/month. Besides the advertised price ($5/account), the vendor offers a deal ($12 for 3 accounts) to the buyer as well. Also, the vendor's Bitcoin address is disclosed in the private messages. As the vendor asks the buyer to vouch for the transaction on the public post, we find the buyer leaves *"Bought from him, all was great :) Account is working"* in the original public thread. Therefore, private messages can reveal much useful information undisclosed in public posts. In the meantime, as shown in Figure 2, private interactions is a crucial step to fulfill the trading in the underground forums. Hence, detecting private interaction can help disclose hidden connections between users, and potentially reveal the goods trading flow.

2.2 Data Overview and Preprocessing

In this paper, we use a dataset from Nulled.io, a very popular hacker forum where users mainly discuss hacking, exploits, and monetization methods. We do not claim this forum represents all underground forums, but it can provide insights into user activities in underground forums. We obtained this dataset from an unknown third party who made it publicly available. Ethically, we do not attempt to identify users in our analyses, and use expressive words to represent corresponding sensitive information, such as "Vendor" and "BTCADDRESS" shown in Figure 1 and Figure 2 to represent seller's user ID and Bitcoin address. Moreover, using leaked and publicly available datasets is an acceptable practice in the study of the underground ecosystem [2, 35].

This dataset has a wealth of information. In particular, it has 599,085 member profiles, including email address, the date of joining the forum, IP address, membership, etc. It has 3,495,596 public posts from Jan 14, 2015, to May 6, 2016, which belong to 121,499 different public post threads. There are also 673,157 user login logs, which contain access time, user ID, and location.

In addition, this dataset has 800,593 private messages, which belong to 404,355 private message threads. These messages have contents, sending time, sender ID, receiver ID, etc. Our preliminary analysis shows that 70.9% of these threads are started by the system or moderators to welcome a new user or to send warning notifications. We exclude those threads from further analysis. As a result, there are 512,227 private messages that belong to 117,708 private message threads and 43,518 user pairs who had private interactions.

To preprocess text data, we retrieved the raw data of public and private messages directly from the forum's database. We used an HTML parser (Beautiful Soup) to remove all these tags. Next, we used a sentence splitter from NLTK (Natural Language Toolkit) to divide the text content into sentences. We also used lemmatizers in NLTK to reduce inflectional forms to a common base form. For example, after this step, "took," "taken," and "takes" will be changed to "take". We also removed all the stop words in the NLTK stop word list from the raw data and punctuation marks from the messages.

3 PRIVATE AND PUBLIC MESSAGE ANALYSIS

In this section, we study the private and public messages from the dumped database for the website Nulled.io. Our goals are: (1) to understand *what* users discuss in private messages, (2) *why* users use private messages, and (3) *how* private message discussions are different from public interactions.

3.1 Content and Purpose

We first investigate the difference between public and private threads in terms of the content and purpose of the communication. Content is the topic the thread is discussing, and purpose is the high-level goal of the thread.

The challenge inherent in answering this question is to recognize the content and purpose for a large number of messages from the database. To solve this challenge, we apply an approach for text topic classification [19, 37]. Specifically, we first explore contents and purposes by manually labeling a subset of the data, and we train a machine learning model to label messages with contents and purposes. To create the manually labeled data, we randomly select 1,000 public and 1,000 private threads from the dump dataset. We then used support vector machines (SVM) for training. To preserve ordering information, we extracted 1-gram, 2-gram, and 3-gram sequences from each thread. We ignored all sequences that appeared in more than 80% of the threads to remove the most frequent ones (as these are used in so many disparate threads and would provide very limited information on the content and purpose). As a result, each thread was represented as a 66,959-dimensional term frequency-inverse document frequency feature vector. We randomly selected 20% of the labeled threads and utilized grid search and five-fold cross-validation to tune the SVM model parameters. We trained separate classifiers for the purpose and content categories. With the optimized parameters, our SVM classifiers have around 0.84 F1 scores on the labeled dataset. In the end, we used the SVM classifiers with tuned parameters to classify all public and private threads.

Table 1: Content and Purpose Categories of Public and Private Threads

	Labeled as	Refined Categories	Public Threads	Private Threads
Content	Bitcoin, Card, Amazon, Money	Monetization	1.7%	0.6%
	Password, Data, Account	Stolen Credential	21.7%	34.6%
	Program, IP, App, Rat, Experience, Hacking, SEO, VPN, Service, Script, Configure Setting, Bot, Website, Proxy, Botnet, Server, Cracking Guide	Hacking-related	48.9%	37.4%
	Video, Story, Tip, Photo, Game, Self Introduction, Device, Rule, Movie, People, Ban, Threads, N/A	Other	27.8%	27.3%
Purpose	Buying, Selling	Trading	13.4%	33.2%
	Sharing	Sharing	48.3%	2.5%
	Help-seeking, Help-offering	Supporting	27.7%	49.5%
	Greeting, Arguing, N/A	Other	10.6%	15.2%

The content and purposes are shown in Table 1. If we cannot label a thread based on the text content, such as a thread with the untranslatable language content, then the interaction purpose and content of this thread will be labeled as *N/A*, We categorize contents to four general classes: `monetization` (e.g., introducing approaches to making money or transferring the money), `stolen credential` (e.g., stolen account credentials, cracked account credentials, or compromised data), `hacking-related` (e.g., hacking services, hacking technique support, and hacking tutorials) and `other`. We also categorize purposes to the following four general classes: `trading` (e.g., selling and buying), `sharing` (e.g., giving away materials or things for free), `supporting` (e.g., seeking or providing help) and `other`.

Table 1 shows the content and purpose statistics in public and private threads. In terms of content, we find hacking, miscellany, and monetization are mentioned more frequently in public than in private, whereas stolen credential is mentioned more frequently in private than in public. This is likely because stolen credential is more valuable, and people would prefer to keep it private. However, monetization is also valuable, yet `monetization` is discussed less in private than in public. Our manual analysis indicates one possible reason of such change is that many of the posters of `monetization` threads request users to contact them through third-party messengers, such as Skype, or directly access their shopping websites.

In terms of purpose, we notice that more public threads aim to share information while more private threads aim to trade, support, greet, or argue, which is aligned with previous findings [39]. Moreover, we find that public sharing is associated with private trading: 96.6% of users who engage in `trading` public threads also have `sharing` public interaction purposes. We believe that this is because a user who publicly shares goods implies that they have more goods, and buyers tend to privately contact the original poster for questions and trading. In a sense, this is a form of advertising. For example, one user often shares many login credentials of cracked game accounts in his/her public posts. Many users privately contact the sharer for trading, such as *"Hey i'd like to buy a lvl 30, i hope it would be possible for me to change the email as well."*

3.2 Artifacts

Based on the content labeling model in Section 3.1, we analyze the popular artifacts in public and private threads. We study login credentials, proxies, contact methods (e.g., email and Skype), and payment methods, and we have the following observations:

Most users constantly keep their personal contacts either in public or in private. For example, among all the email contacts found in private threads, only 1.9% of the email contacts appear in both public and private threads. Also, users tend to keep their Skype ID secret until the end of the trading. We notice that only 41.1% of Skype-mentioned private threads include Skype IDs.The analysis shows that many users who engage in Skype-mentioned private threads cannot reach initial deals before exchanging Skype IDs or they prefer to use the built-in private messaging function of the forum.

In addition, we notice that privately exchanged goods are also disclosed in public. For example, only 64.2% and 40.2% of privately exchanged credentials and proxies, which are most frequently exchanged in public, are never discussed in public.

Forum administrators also host multiple payment accounts. Those accounts are used to collect the administrative fee such as account upgrading and tested products purchasing. We found 7 payment accounts, and 6 of them are found from private threads. This implies that one should establish private interaction with the administrators to investigate the financial status of the underground forum.

Previous research has conducted per-user analysis, based on the fact that duplicated users are typically not allowed in the underground forum. However, we observe that 37 out of 1,165 PayPal accounts are used by multiple users, which implies that duplicated users exist in the underground forum. Our manual analysis indicates that these forum accounts likely to belong to the same person, because these users are performing the same business, clearly leaving the same private contact information, such as Skype IDs. Also, we notice that these users, who have multiple accounts linked by the same PayPal account, have used precautions to avoid being identified by the forum's Duplicated-Account detection system: they use different IP addresses when logging into their different user accounts. They also take care to use email addresses and usernames that are very different from their other accounts. Additionally, we notice that nearly 86% of such multiple-linked accounts are identified from the private threads, therefore demonstrating the importance of private messages to understanding an underground marketplace.

3.3 Private Message Recipients Behavior

A *private message recipient* is a user that is initially contacted in a private thread. As private messages can contain more security sensitive contents and purposes, studying private message recipients will help us understand how private interactions are initiated. In this subsection, we study two questions:

- What is the difference in the public activities of private message recipients and other users? Specifically, what is the difference in public security messages, posts, and replies, and private security posts and replies?
- How do private message recipients behave in public as compared to in private?

Table 2: Number of Recipients in Different Public and Private Interaction Categories

		# Users in Public	# Recipients in Public	# Recipients in Private
Content	Monetization (a)	25,448 (4.2%)	6,524 (38.4%)	26 (0.2%)
	Stolen Credential (b)	93,645 (15.6%)	13,571 (79.8%)	6,817 (40.1%)
	Hacking-related (c)	279,148 (46.6%)	16,481 (97.0%)	7,221 (42.5%)
	$a \wedge b$	18,679 (3.1%)	6,222 (36.6%)	22 (0.1%)
	$a \wedge c$	24,509 (4.1%)	6,498 (38.2%)	23 (0.1%)
	$b \wedge c$	81,824 (13.7%)	13,387(78.8%)	5,096 (30.0%)
	$a \wedge b \wedge c$	18,463 (3.1%)	6,203 (36.5%)	21 (0.1%)
	Total	291,692	16,672	8,944
Purpose	Trading (d)	32,988 (5.5%)	9,823 (57.8%)	6,983 (41.1%)
	Sharing (e)	288,403 (48.1%)	16,501 (97.1%)	541 (3.2%)
	Supporting (f)	92,091 (15.4%)	14,224 (83.7%)	8,151 (48.0%)
	$d \wedge e$	31,867 (5.3%)	9,697 (57.1%)	471 (2.8%)
	$d \wedge f$	24,021 (4.0%)	8,946 (52.6%)	5,629 (33.1%)
	$e \wedge f$	84,121 (14.0%)	14,090 (82.9%)	495 (2.9%)
	$d \wedge e \wedge f$	23,837 (4.0%)	8,916 (52.5%)	450 (2.6%)
	Total	297,310	16,731	9,530

To answer these questions, we take advantage of messages with content and purpose labels (§ 3.1). We divide the messages into two groups by whether its author is a private message recipient or not. For each group, we calculate the proportion of users involved in the messages with different content and different purposes. In our dataset, there are 43,518 user pairs who have private interactions, and 16,997 users are `recipients`. The statistic results of this analysis are shown in Table 2.

What is the difference in the public activities of private message recipients and other users? We study the differences in three different aspects: (1) security messages, (2) content-specific and purpose-specific messages, and (3) posts and replies. Security messages are messages with security-related content or purposes. As we stated in Table 1, in this paper we consider monetization, stolen credential, and hacking as security-related contents, and we consider trading, sharing, and supporting as security-related purposes. Content-specific and purpose-specific messages are labeled with one or multiple content or purposes from the above categories. Posts and replies are two types of messages in public thread. Posts are the messages initializing new threads, while replies are messages posted as follow-ups to existing threads.

- **Security messages.** Column # Recipients in Public and column # Users in Public shows the difference of public activities between private message recipients and other users. Private message recipients are more involved in security-related messages. For example, 98.1% of the private message recipients have discussed monetization, stolen credential, or hacking-related topics, whereas 48.6% of the other users have discussed this content. This difference also implies that users involved in more security-related activities will be more likely to be contacted privately.
- **Content-specific and purpose-specific messages.** In particular, we study the difference in posts and replies with specific contents and purposes. Table 2 shows the number of unique private message recipients that post/reply a message in a specific topic or purpose, as well as the number of unique recipients

that are contacted for the same topic and purpose. For stolen credential, hacking, trading, and supporting, more than 50% of the recipients are contacted with the same content/purpose after posting. However, for monetization and sharing, few recipients are contacted after posting.

- **Posts and replies.** We compare the total number of the unique authors to that of unique private message recipients, as shown in Table 3. Based on these results, we observe that a majority of post authors are also private message recipients. For example, 906 out of 1,296 (69.9%) authors that post monetization messages become recipients. However, many fewer authors are contacted due to their replies. Among 24,152 unique authors that reply to monetization messages, only 5,618 (23.2%) are recipients that are privately contacted by other users. This implies that to attract the other users for private interaction, one should post security messages rather than reply to existing threads.

In addition, we notice from Table 3 that there are more `repliers` than `posters` in all the categories. Our results show that most of `posters` never reply in his/her created thread, which indicates that `repliers` cannot receive any help from `posters` by replying to their posts. Especially, many initial posts explicitly request users to privately contact the post's creator. Because of this results, we are interested in the reasons for publicly *replying* to posts. Our manual analysis indicates one possible reason: `posters` attempts to have a lower *leecher* value. Because many users attempt to take advantage of the underground forum resources without making contributions, the underground forum assigns a *leecher* value to each user, which is used to quantify a user's contribution, where a lower value is better. This value is rated by the system automatically based on various metrics, such as the number of threads a user created, the number of replies a user obtained, etc. If a user has a high *leecher* value, then many of his/her activities will be restricted, such as private message limits, being unable to access particular types of post, etc. To lower the *leecher* value, a user needs his/her posts to have more replies, which imply that more users are interested in his/her posts. Therefore, most of the `posters` use a feature of the forum to hide essential content from the post, and this content is only revealed when other users reply to the post. This also explains why the text content of most of the replies are meaningless, such as *"thx", "ty"*.

How do private message recipients behave in public as compared to in private? Column # Recipients in Public and column # Recipients in Private show the difference of private message recipients' activities between public messages and private messages. Interestingly, we find that private message recipients discuss *less* about security in private than in public. While 98.5% of the recipients message publicly for trading, sharing, and supporting, only 56.1% of them message privately with similar purposes. Instead, these recipients use private threads for other purposes, such as to argue for reviews in public threads.

Posting is more likely than replying in having similar content or purpose in both public and private interactions. As shown in Table 4, column # Recipient-posterin Public and column # Recipient-posterin Both Public and Private shows the difference of private message `recipient-posters`' activities between public and private messages. In general, 6,807 out of 10,205 (66.7%) `recipient-posters` have similar security related content in both public and private, while it is only 32.5% for `recipient-repliers`.

Table 3: Poster and Replier Statistics Based on Their Public Thread Contents and Interaction Purposes

		# Recipient-posters	# Posters	# Recipient-repliers	# Repliers
Content	Monetization (*a*)	906	1,296	5,618	24,152
	Stolen Credential (*b*)	5,277	8,491	8,294	85,154
	Hacking-related (*c*)	8,576	21,172	7,905	257,976
	a ∧ b	631	699	2,888	14,219
	a ∧ c	735	914	2,065	17,087
	b ∧ c	3,739	4,937	4,424	63,666
	a ∧ b ∧ c	551	596	1,339	10,625
Purpose	Trading (*d*)	4,100	6,093	5,723	26,895
	Sharing (*e*)	7,741	17,684	8,760	270,719
	Supporting (*f*)	6,609	16,558	7,615	75,533
	d ∧ e	2,685	3,260	2,895	20,671
	d ∧ f	2,500	3,112	3,035	14,404
	e ∧ f	4,367	6,846	4,763	61,429
	d ∧ e ∧ f	1,935	2,215	1,849	11,696

Table 4: Number of Recipient-posters and Recipient-repliers in Different Public and Private Interaction Categories

		# Recipient-posters in Public	# Recipient-posters in Both Public and Private	# Recipient-repliers in Public	# Recipient-repliers in Both Public and Private
Content	Monetization (*a*)	906	6	5,618	14
	Stolen Credential (*b*)	5,277	3,397	8,294	2,900
	Hacking-related (*c*)	8,576	4,836	7,905	2,301
	a ∧ b	631	5	2,888	5
	a ∧ c	735	6	2,065	4
	b ∧ c	3,739	2,287	4,424	732
	a ∧ b ∧ c	551	5	1,339	3
	Total	10,205	6,807	13,779	4,477
Purpose	Trading (*d*)	4,100	2,860	5,723	2,557
	Sharing (*e*)	7,741	405	8,760	133
	Supporting (*f*)	6,609	4,454	7,615	3,088
	d ∧ e	2,685	269	2,895	43
	d ∧ f	2,500	1,863	3,035	862
	e ∧ f	4,367	313	4,763	47
	d ∧ e ∧ f	1,935	228	1,849	21
	Total	10,833	7,464	13,254	4,847

4 DETECTING PRIVATE INTERACTIONS IN UNDERGROUND FORUMS

In this section, we study the *detection* of private interactions by using two users' publicly available information in underground forums. Because users trade illicit goods or services through private interactions, detection of these private interactions helps to identify illegal activities, trace goods flows, and disclose hidden connections.

4.1 Approach Overview

We apply machine learning methods to detect private interactions. We first train machine learning models with constructed user pair instances. Then, we apply the trained models on the testing data and evaluate based on precision, recall, F1 score, and accuracy.

Data Selection. In the `Nulled.io` dataset, there are much more user pairs that do not have private interactions than whose who have. For all of the 599,085 users, only 43,518 user pairs ever have private interactions. We randomly sub-sample user pairs of non-private interaction to avoid the common problems of an imbalanced dataset, such as the bias towards the majority class [10]. Also, we adopt data pruning and cross-field validation in sub-sampling five times to train each classifier.

Feature Extraction. Based on our observations (described in previous sections), we synthesize three categories of features for the private interaction detection: (1) features from a user's profile, (2) features from a user's public activities, and (3) features from a user pair, as shown in Table 5. The features are from *publicly available and objective information* of a user pair. We choose these features under the following considerations. First, as users try to hide their real identities in the underground forum, they may provide fake or incomplete information. Second, unlike public social networks, underground forums only need users to provide a small amount of profile information, and much of the information is hidden from the public. Also, only a small percentage of users have private interactions, and users typically have a different focus when privately interaction compared to their public interactions.

Training. We use Naive Bayesian (NB), Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), AdaBoost, Multi-Layer Perceptron (MLP), and Random Forest (RF) to detect private interactions in the underground forum. Table 6 shows the best performance configurations of each classifier.

4.2 Testing Results

4.2.1 Detection Performance of Private Interactions of Overall Users. We used five-fold cross-validation in our experiments, and all features were normalized. Table 7 shows the detection performance of each classifier. The precision is the fraction of correctly identified private interaction instances among all of the detected private interaction instances. The recall indicates the percentage of private interaction instances are correctly identified. The F1 score is used to measure the overall performance while considering both the precision and recall.

As shown in Table 7, the algorithms can effectively detect private interactions. Most of the algorithms can achieve higher than 0.85 in precision and 0.92 in the recall. In particular, ensemble algorithms, AdaBoost, and RF, and the neural network algorithm, MLP, have the highest F1 score. In the meantime, MLP has the best performance on precision. Additionally, RF outperforms other algorithms on overall accuracy and recall. As AdaBoost, MLP, and RF have at least one of the best measurement performances, we focus on these three algorithms in the rest of the paper.

Figure 3 depicts the ROC curves of detecting private interactions by using the three algorithms. In general, all of the algorithms have a high true positive rate (TPR) and low false positive rate (FPR). In particular, RF and MLP outperform Adaboost. Additionally, we observe that, with 10% FPR, MLP can achieve more than 95.5% TPR and RF has less than 4% false negative rate (FNR).

4.2.2 Detection Performance of Private Interactions of Publicly Active Users. As shown in Figure 4, the top 10% of privately active users engage in 80% of the total private interactions, and users who are in the top 10% publicly active engage in nearly 90% of total private interactions. To detect private interactions of top publicly active users, we constructed multiple experimental subsets from the dataset containing private interaction user pairs that consisted of users who were top n of publicly active (where n is set to 5%, 10%, and 15%). We used the same number of user pairs of non-private interactions and user pairs with private interactions to form balanced subsets.

Table 8 shows the detection performance results using AdaBoost, MLP, and RF. "PI Coverage" in Table 8 indicates the percentage of private interactions involving the users who are in the top n% of publicly active users. Note that 79.2% of private interactions attribute to the top 5% of publicly active users. As clearly shown in Table 8, the classifiers have better performance in detecting private interactions for the top n% of publicly active users.

Figure 5 shows the ROC curves of detecting private interactions of users pairs that contain the top 5% of publicly active users. Different from overall detection performance, Adaboost has a better performance than MLP. Moreover, the algorithms have a higher TPR in Figure 5 than in Figure 3 at the same FPR.

5 ADVERSARIAL ANALYSIS

In Section 4, we show that our approach can effectively detect private interactions in the underground forum. In this section, we perform adversarial attacks analysis to evaluate the robustness of our detection technique.

Because many users prefer to discuss and exchange illicit detailed information in private, in order to escape detections they may *intentionally change their public behaviors* in the underground forums. Also, to maintain the underground forums normal operations, to protect and attract users, administrators of underground forums could perform actions to prevent their users' private interactions from being detected. Therefore, it is crucial to evaluate the robustness of our detector in different adversarial scenarios. Because the adversarial attack purpose is to escape the detection, we only consider users pairs who have private interactions and measure the accuracy of successfully detected private interaction given different adversarial attack scenarios.

In this section, we analyze two scenarios: (1) The *evasion attack*, where users in underground forums adjust their behaviors and hide in other users to escape detection, and (2) The *poisoning attack*, where administrators of underground forums generate fake avatars and manipulate their activities to poison the data to prevent their users' private interactions from being detected.

5.1 Evasion Attack

Evasion attacks refer to adversaries in underground forums that adjust their behaviors to hide in the crowd to escape detection. By analyzing the features we used in the detector, we notice that some of them can be modified if adversaries intentionally change their behaviors in the underground forum. We have marked each feature as changeable, unchangeable, or partial changeable in Table 5. *Membership* and *Banned* are partially changeable features because their statuses are not fully controlled by the users. For example, a user can change his/her behaviors in the underground forum to be banned, but recover his unbanned status from banned is decided by the administrators. Also, *Leecher* in the *Reputation* can be changed by adversaries depending on how much the contribution the user makes to the underground forum. However, *Reputation* and *Like* are rated by other people that are not controlled by the adversary.

Table 5: Features Used in the Detector

	Feature Name	Explanation
Profile Features	Membership [p]	To create a generic approach, we categorize memberships into five classes: (1) Basic membership, such as *Member*; (2) Upgraded membership, such as *Royal*; (3) Limited membership, such as *Banned*; (4) Fee-based membership, such as *VIP*; (5) Staff, such as *Administrator*.
	JoinDate [u]	The time of joining the forum
	LastVisit [u]	The time of the last visit
	Views [u]	The number of views on the user's profile. A user's profile includes user's basic information, reputation values, and activity history.
	Reputation [c,u]	Many forums use multiple types of reputation values to show a user's contributions, honors, and trustworthiness. For example, the Nulled.io forum uses three such values, which are: leecher, like, and reputation, to indicate the user's contributions, honors, and trustworthiness, respectively.
	Banned [p]	A user's status: This is used to show if a user's account is banned in the forum. When a user violates the forum policy, such as creating duplicated accounts and spamming, then the staff members may ban his/her account.
Activity Features	Posts [c]	The number of a user's public posts.
	Threads [c]	The number of public threads a user engages in. This feature shows a user's public activities on different public threads.
	Topics [c]	The number of public threads initiated by a user. This feature indicates how often a user opens new public post topics in an underground forum.
	ThreadViews [u]	The number of views on a user's public threads. This feature indicates the popularity of a user's public threads.
	Subforums [c]	The number of subforums a user is involved in. The underground forum has several subforums for different themes, and the post threads must be published in the corresponding subforum. This feature indicates the number of general themes a user is interested in.
	Friends [c]	The number of users who publicly interact with the user. This feature shows the degree centrality of a user in an underground forum network, which is formed by users' public interactions.
Interaction Features	Interactions [c]	The number of this user pair's public interactions.
	CommonTopics [c]	he number of public threads that have both users' posts. This feature shows how often these two users are interested in the same public post topics.
	CommonSubforums [c]	The number of subforums that both users are involved in. This feature indicates how often these two users are interested in the same general themes in an underground forum.
	CommonFriends [c]	The number of users who publicly interact with both users. This feature reveals the number of direct neighbors shared by these two users in an underground forum interaction network.

[c] Changeable Feature
[p] Partial Changeable Feature
[u] Unchangeable Feature

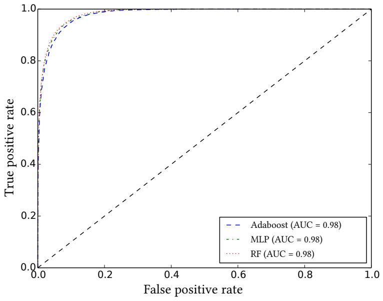


Figure 3: ROC curves of detection of overall user pairs

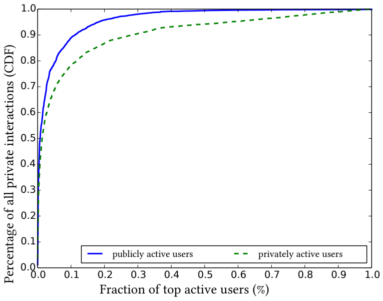


Figure 4: The growth of private interactions over the fraction of top active users

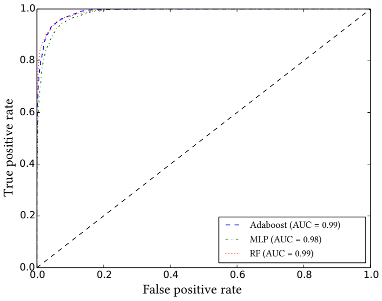


Figure 5: ROC curves of detection of user pairs that contain top 5% of publicly active users

Table 6: Classifier Configurations

Algorithm	Configurations
NB	alpha = 0.001; fit_prior = "True"
LR	C = 10; solver = "liblinear"
SVM	kernel = "rbf"; C = 100; gamma = 0.01
KNN	n_neighbors = 11; weights = "distance"; algorithm = "ball_tree"
AdaBoost	n_estimators = 250; learning_rate = 1; algorithm = "SAMME.R"
MLP	activation = "logistic"; solver = "adam"; learning_rate = "adaptive"; hidden_layer_sizes = (150,)
RF	n_estimators = 250; max_features = "sqrt"

Table 7: Detection Performance in the Underground Forum

Algorithm	Precision	Recall	F1 Score	Accuracy
NB	0.85	0.80	0.82	0.82
LR	0.85	0.92	0.88	0.88
SVM	0.90	0.93	0.92	0.91
KNN	0.87	0.94	0.90	0.89
AdaBoost	0.91	0.94	**0.93**	0.92
MLP	**0.93**	0.94	**0.93**	0.93
RF	0.92	**0.96**	**0.93**	**0.94**

In the evasion attack, we focus on two main situations: adversaries modify their changeable features without a strategy and with a strategy. We leave as future work the more delicate adversarial machine learning schemes on our model, as well as the assessment of the robustness of the detector against more sophisticated data evasion attacks.

5.1.1 *Non-strategic Attack.*

In the non-strategic attack, adversaries do not know what features are used in the detector, and they attempt to assign random values to all changeable features by adjusting their activities. In this situation, we consider two kinds of attacks: (1) single-user attack, where one user does not want his/her private interactions to be detected, and (2) two-user attack, where two users try to prevent their private interactions from being detected.

Single-user Non-strategic Attack. In this attack, the adversary does not consider a specific private contact user, so this user only

Table 8: Detection Performance of User Pairs Containing Top *n*% of Publicly Active Users

Top N	Performance	AdaBoost	MLP	RF	PI Coverage
	Precision	**0.94**	0.93	0.92	
5%	Recall	0.97	0.95	**0.98**	79.2%
	F1 Score	**0.95**	0.94	**0.95**	
	Precision	**0.94**	0.93	0.93	
10%	Recall	0.96	0.95	**0.97**	88.9%
	F1 Score	**0.95**	0.94	**0.95**	
	Precision	**0.93**	0.92	0.92	
15%	Recall	0.96	0.94	**0.97**	93.1%
	F1 Score	**0.95**	0.93	**0.95**	

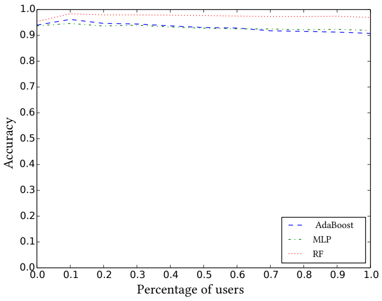


Figure 6: Detector performance on the single-user non-strategic attack

modifies his/her changeable profile and activity features to random values. To evaluate the robustness of the detector to this attack, we randomize the value of each changeable feature within a reasonable scope, which are determined by sampling from a subset of users.

It is worth pointing out that changing an adversary's activity feature values can influence his/her interaction feature values with other users. For example, the number of common topics between two users is impacted if one of the users changes his/her *Topics* value. Because the user can delete posts from topics and make new posts in new topics, or just delete or make new posts, it is difficult to know how the change impacts the number of common topics between the users. Therefore, we use Equation 1 to adjust the interaction feature values in each of the instances.

$$InteractionVal' = ChangeableVal' \times \frac{InteractionVal}{ChangeableVal} \tag{1}$$

In this example, *InteractionVal'* is the adjusted number of *CommonTopics*. *ChangeableVal'* is the randomly generated number of *Topics*. *InteractionVal* is the original number of *CommonTopics*, and *ChangeableVal* is the original number of *Topics*. To have fair results, we perform the evaluation 1,000 times, and the average results are shown in Figure 6.

Figure 6 shows the performance of private interaction detection when an adversary modifies his/her changeable information without a strategy by observing differing percentages of other users. As shown in Figure 6, our detector is robust to this kind of attack and RF outperforms other algorithms. We notice that the successful evasion rate decreases when considering a larger percentage of users.

Two-user Non-strategic Attack. In this attack, two adversaries have clear private contact targets and try to avoid detection of their private interactions. Therefore, there are three methods to perform this attack: (1) both adversaries change their changeable features in their profiles and activities, and keep the interaction features the same, (2) both adversaries modify their changeable features, and (3) both adversaries only modify their changeable interaction features. To simulate this attack, we randomize the value within a reasonable scope by sampling from a subset of users. Note that interaction features and their corresponding activity features have a relationship that constrains the randomly generated values—e.g., in the first method of this attack, as *CommonTopics* value is not changed, both adversaries' new *Topics* values should be between *CommonTopics* and maximum *Topics* value of a subset of users.

Figure 7 shows the detector performance on the two-user non-strategic attack with different methods by sampling different percentage of users. As shown in Figure 7, the detector is robust against the two-user non-strategic attack, with more than 80% accuracy in general. Also, RF outperforms other algorithms in all methods; modifying the interaction feature values can impact the RF performance. Additionally, in the two-user attack, adversaries only need to consider 10% of users' information to obtain their considerable evasion rate.

In summary: the detector is robust to the non-strategic evasion attacks, and two-user non-strategic attacks are more challenging to the detector than single-user non-strategic attacks. In addition, RF always outperforms other algorithms in our evaluations.

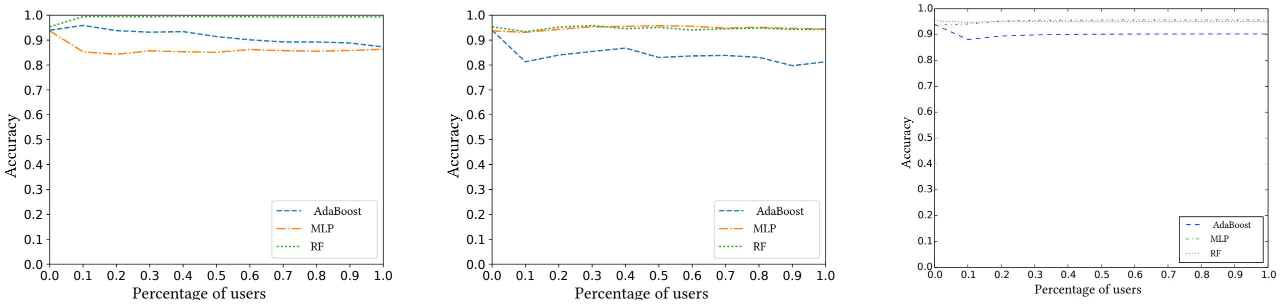
5.1.2 *Strategic Attack.*

Since many publications discuss machine learning based detectors, such as [3, 6, 27], and the publicly available information in the underground forum is limited, it is possible for adversaries to guess the potential features used in the detector. Therefore, to hide themselves in the crowd, adversaries are motivated to assign specific values to a minimum number of features. In this section, we assume adversaries already know the changeable features and the value scope of each feature. In this situation, we still consider the aforementioned two kinds of attacks: (1) single-user attack and (2) two-user attack.

Single-user Strategic Attack. In this attack, an adversary does not consider a specific private contact user, so this adversary only modifies changeable features in his/her profile and activity. To evaluate the robustness of the detector, an adversary only assigns a specific value to one feature at a time. We also adopt the same approach to adjust interaction feature values (Equation 1).

Figure 8 shows the detector performance when a user performs a single-user strategic attack in terms of different algorithms. The x-axis shows within a reasonable value range of a feature, the value that will be assigned to the feature. The y-axis indicates the accuracy of successfully detected private interactions. As shown in Figure 8, the detector with RF has the most robustness (the lowest accuracy is 83.3%) to the single-user strategic attack and the MLP is the least robust one with only 32.2% accuracy. Additionally, *Friends, Topics,* and *Leecher* impact the detector more significantly than other features. Compared with the single-user non-strategic attack, this attack can increase the adversaries chance to evade detection.

Two-user Strategic Attack. In this attack, adversaries assign a specific value to one type of their changeable features at a time.

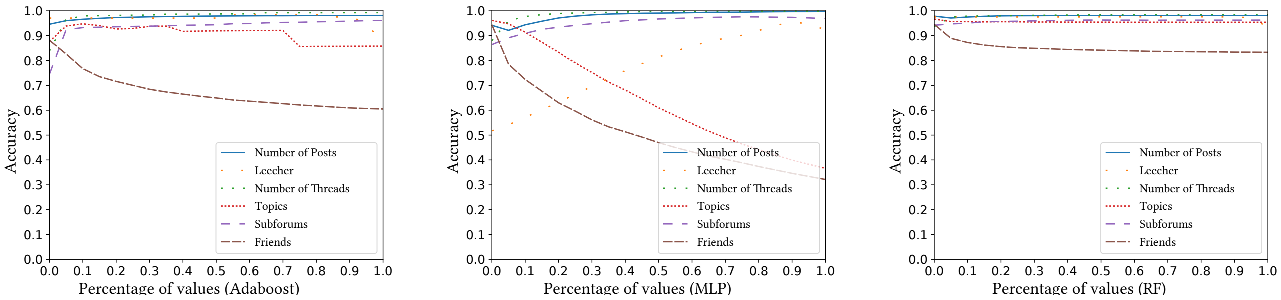


(a)

(b)

(c)

Figure 7: Detector performance on different two-user non-strategic attack methods



(a)

(b)

(c)

Figure 8: Detector performance on the single-user strategic attack

For example, both users modify their *Topics* features, while keeping other feature values unchanged. As explained previously, interaction features and their corresponding activity features have a relationship that constrains the value scope.

Figure 9 shows the detector performance when two adversaries perform a strategic attack in terms of different algorithms. The x-axis shows within a reasonable value range of a feature, the value that will be assigned to the feature. The y-axis indicates the accuracy of successfully detected private interactions. As shown in Figure 9, RF is the most robust algorithm in the detector, and its lowest accuracy is 91.2%. In contrast, MLP has the worst performance that is 3.9% accuracy. Also, to our surprise, interaction features do not impact the detection performance significantly, as the lowest accuracy are obtained when modifying *Friends, Topic,* and *Leecher* in Adaboost and MLP based detectors, as shown in Figure 9(a) and Figure 9(b). Although *Interactions* is more influential than other features in Figure 9(c), the detector still has more than 90% accuracy. Comparing this attack with the single-user strategic attack, although the two-user strategic attack has the lowest accuracy in MLP, frequently the single user strategic attack is more challenging to our detector.

In summary: strategic attacks have a higher chance to evade detection than non-strategic attacks. Additionally, single-user attacks are more challenging than two-user attacks in the strategic attacks. Also, interaction features are less influential than profile and activity features.

Table 9: Detection Performance with Unchangeable Features

Algorithm	Precision	Recall	F1 Score	Accuracy
AdaBoost	0.84	0.86	0.85	0.85
MLP	0.84	0.91	0.88	0.87
RF	**0.85**	**0.94**	**0.89**	**0.89**

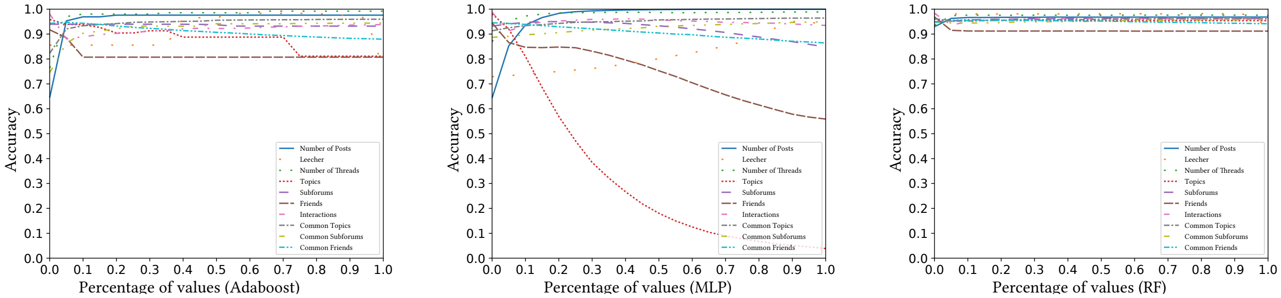
5.1.3 *Counter Evasion Attack.*

To counter the evasion attack, we evaluate our detector robustness with *only using unchangeable features*. The new evaluation indicates that the detector still has considerable performance, shown in Table 9.

Figure 10 depicts the ROC curves of detecting private interactions by using unchangeable features. In general, using unchangeable features has worse performance than using all features by comparing this ROC curves with Figure 3 and Figure 5. In particular, all algorithms can have more than 0.90 TPR with 0.18 FPR. Also, RF outperforms another two algorithms.

5.2 Poisoning Attack

To make underground forums operate properly, attract more users, and protect their users' privacies, administrators of such underground forums need to take actions to help prevent their users'



(a)

(b)

(c)

Figure 9: Detector performance on the two-user strategic attack

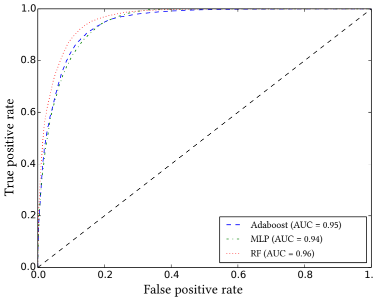


Figure 10: ROC curves of detection by using unchangeable features

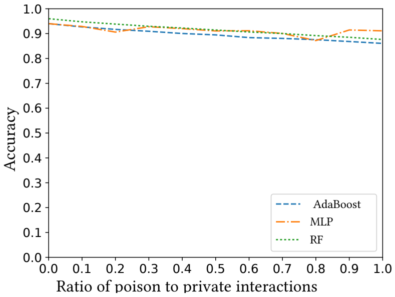


Figure 11: Detector performance on the random poisoning attack

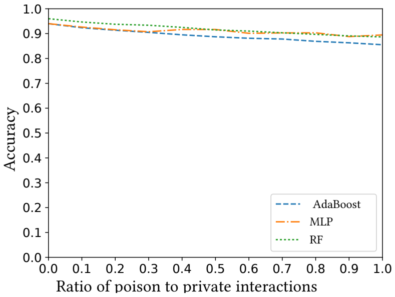


Figure 12: Detector performance on the normal-user-based poisoning attack

private interactions from being detected. As administrators have fully control of the underground forums, they can generate fake avatars and manipulate their activities to make false private interaction samples to pollute the data.

By considering the approach to create fake avatars, there are two types of poisoning attacks: (1) the *random poisoning attack*, where all individual feature values of the fake avatars are generated randomly within a reasonable scope based on all users' information, and interaction feature values of fake avatars are generated based on their individual features. (2) The *normal-user-based poisoning attack*, where administrators find out all normal users, who never have private interactions from the database, and create fake avatars by using normal users' information: while constructing the false samples, the interaction feature values are randomly generated that are within reasonable scope based on two randomly selected fake avatars.

Note that we *do not* construct false samples between fake avatars and real users while considering practical situations. Letting a fake avatar privately contact a real user will provide a bad experience to the user because it means that the user has strange private messages with unknown people. This could lead the user to feel that his/her account is stolen.

To evaluate the robustness of the detector to this attack, we first split the clean dataset into training and testing datasets. Secondly, we generate different numbers of false private interaction samples and then mix them with real training dataset. Finally, we build our detector based on the polluted training dataset and apply the testing dataset to evaluate the detector performance.

Figure 11 and Figure 12 show the detector performance when administrators perform two types of poisoning attacks by injecting different numbers of false private interaction samples. The x-axis shows the ratio of false samples to the real private interactions, and the y-axis indicates the accuracy of successfully detected private interactions in the testing dataset.

As shown in Figure 11 and Figure 12, the detector is robust to both types of poisoning attacks, and the performances are very similar. In general, MLP outperforms other algorithms in the poisoning attack with the lowest 0.91 accuracy in random poisoning attack and 0.89 accuracy in normal-user-based poisoning attack. Additionally, injecting more false private interaction samples will decrease the detector performance, but the influence is limited. In the meantime, by comparing the detection performance in the different poisoning attacks, the normal-user-based poisoning attack is more challenging to the detector.

6 DISCUSSION

In this section, we explore the limitations of the data analysis and private interaction detection approaches, and we present the direction of our future work.

In this paper, we use Nulled.io as the subject in our study. Although Nulled.io is one of the most popular underground forums and can be representative for the security research for private interactions, a single forum may be insufficient for data analysis and private interaction detection. We will look for and investigate more data for future work.

For data analysis, we group messages into different categories in terms of contents and purposes. The categories are generated manually based on 1000 samples of public and private message in respective. Due to the limit of samples, the categories may not be complete. In the future, we will look into approaches towards complete categories for both content and purposes. Moreover, we label messages by applying SVM with supervised classification. However, the SVM model may be imprecise in labeling messages. Data labeling is known to be open in data analysis research. We leave more accurate labeling as future work.

For private interaction detection, we test the robustness of the detection model against adversarial machine learning attacks such as data evasion and data poisoning attacks. For both types of the attack, we consider hard-coded strategies for data generation. One future direction is to test our detection model on more sophisticated adversarial machine learning techniques [5, 12, 48]. Also, we will explore more robust machine learning models for private interaction detection, such as to apply the machine unlearning technique [9] to the current model.

7 RELATED WORK

Analysis and measurement of underground forums and marketplaces. Research efforts have been invested in understanding the organizational structure and social activities of underground forums and marketplaces [2, 35, 46]. Motoyama et al. studied six underground forums for understanding what the products and services were exchanged [35]. Also, researchers tried to identify anonymous authors of texts in underground forums by analyzing their writing styles [2]. Zhao et al. analyzed social dynamics relevant to net-centric attacks to discover adversarial evidence [46]. Additionally, Hao et al. analyzed the reshipping service from underground forums to show how cybercriminals monetize stolen credit cards and the relationships between different actors who are involved in this scam activities [20]. Also, Thomas et al. investigated web services that create and trade fraudulent accounts by cybercriminals in underground markets [44]. Even though Radianti et al. pointed out that users in underground forums and marketplaces prefer to discuss details in private communication channels [39], private messages in the underground society have been largely overlooked.

Messaging in online social networks. People like to post diverse types of message to share their status, moods, opinions, etc., in online social networks. The motivations and purposes of messaging in online social networks have been studied [25, 29, 34]. In the meantime, researchers also investigate how the personality and motivations related to the communications in online social networks [40, 42]. Also, many works are conducted to analyze the text contents from different aspects, such as extracting and categorizing topics [28, 34, 36], analyzing the text to discover the social structure [15, 32] and showing users' sentiments [22, 24, 30, 38]. However, in underground forums, the messages are full of leetspeak, cyber jargons, and users' motivations are different from using public online social networks. In the meantime, users prefer to show their actual purposes in private messages instead of public posts.

Link prediction in online social networks. Link prediction, which has been widely used to suggest friends in online social networks [4, 11, 14, 27, 41, 45], inspires us to design our private interaction detection algorithms because the private interaction is a kind of hidden link between users in the underground forum. Even though they share similarities, link prediction in public social networks and private interaction detection in underground forums are fundamentally different: 1) most link prediction approaches are based on network topology analysis [16]. They assume if two users have shared friends it is highly possible they have or should have a direct connection. However, in underground forums, users do their best to hide their real identities. Hence, the connection between users in the underground forum is not based on the real-world identities, but the anonymous public interactions and 2) in public social networks, users usually use their real identities and profile information [18]. Nevertheless, the self-provided information in an underground forum may not be trustworthy and cannot be used for the detection [23].

8 CONCLUSION

Analyzing underground forums and marketplaces is of great importance to understand and combat cybercrime and illegal activities. Even though research efforts have been invested in understanding the organizational structure and social activities of underground forums, private messages have been overlooked. In this paper, we analyze an underground forum Nulled.io to understand what users discuss in private messages, why users are contacted privately, etc. In addition, we designed machine learning models that take the characteristics of underground forums into account to detect private interactions between users. The results showed that our models are effective in detecting private interactions and can withstand attacks.

ACKNOWLEDGMENTS

This work was supported in part by grants from the U.S. Army Research Laboratory and the Center for Cybersecurity and Digital Forensics at Arizona State University. The information reported here does not reflect the position or the policy of the funding agency or project sponsor.

REFERENCES

[1] Sadia Afroz, Vaibhav Garg, Damon McCoy, and Rachel Greenstadt. 2013. Honor among thieves: A common's analysis of cybercrime economies. In *eCrime Researchers Summit (eCRS)*.

[2] Sadia Afroz, Aylin Caliskan Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of the IEEE Symposium on Security and Privacy*.

[3] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

[4] Mohammad Al Hasan and Mohammed J Zaki. 2011. A survey of link prediction in social networks. In *Social network data analytics*.

[5] Mostafa D Awheda and Howard M Schwartz. 2016. A fuzzy reinforcement learning algorithm using a predictor for pursuit-evasion games. In *Proceedings of the IEEE International Systems Conference (SysCon)*.

[6] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. 2010. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*.

[7] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. 2009. Characterizing user behavior in online social networks. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*.

[8] Moira Burke and Robert E Kraut. 2014. Growing closer on facebook: changes in tie strength through social network site use. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[9] Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy*.

[10] Nitesh V Chawla. 2009. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*.

[11] Hsinchun Chen, Xin Li, and Zan Huang. 2005. Link prediction approach to collaborative filtering. In *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*.

[12] Lingwei Chen, Yanfang Ye, and Thirimachos Bourlai. 2017. Adversarial Machine Learning in Malware Detection: Arms Race between Evasion Attack and Defense. In *Proceedings of the IEEE European Intelligence and Security Informatics Conference (EISIC)*.

[13] Nicolas Christin. 2013. Traveling the Silk Road: A measurement analysis of a large anonymous online marketplace. In *Proceedings of the International World Wide Web Conference (WWW)*.

[14] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* (2008).

[15] Jana Diesner and Kathleen M Carley. 2005. Revealing social structure from texts: meta-matrix text analysis as a novel method for network text analysis. In *Causal mapping for research in information technology*.

[16] Yuxiao Dong, Jie Tang, Sen Wu, Jilei Tian, Nitesh V Chawla, Jinghai Rao, and Huanhuan Cao. 2012. Link prediction and recommendation across heterogeneous social networks. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*.

[17] Greg Durrett, Jonathan K Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca S Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Identifying Products in Online Cybercrime Marketplaces: A Dataset for Fine-grained Domain Adaptation. *arXiv preprint arXiv:1708.09609* (2017).

[18] Nicole B Ellison, Charles Steinfield, and Cliff Lampe. 2007. The benefits of Facebook "friends:" Social capital and college students' use of online social network sites. *Journal of computer-mediated communication* (2007).

[19] Jason Franklin, Adrian Perrig, Vern Paxson, and Stefan Savage. 2007. An inquiry into the nature and causes of the wealth of internet miscreants.. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

[20] Shuang Hao, Kevin Borgolte, Nick Nikiforakis, Gianluca Stringhini, Manuel Egele, Michael Eubanks, Brian Krebs, and Giovanni Vigna. 2015. Drops for stuff: An analysis of reshipping mule scams. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*.

[21] Thorsten Holz, Markus Engelberth, and Felix Freiling. 2009. Learning more about the underground economy: A case-study of keyloggers and dropzones. In *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*.

[22] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu. 2013. Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*.

[23] Haruna Isah, Daniel Neagu, and Paul Trundle. 2015. Bipartite network model for inferring hidden ties in crime data. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.

[24] Aamera ZH Khan, Mohammad Atique, and VM Thakare. 2015. Combining lexicon-based and learning-based methods for Twitter sentiment analysis. *International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE)* (2015).

[25] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media?. In *Proceedings of the International World Wide Web Conference (WWW)*.

[26] Cliff AC Lampe, Nicole Ellison, and Charles Steinfield. 2007. A familiar face (book): profile elements as signals in an online social network. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[27] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology* (2007).

[28] Kar Wai Lim, Changyou Chen, and Wray Buntine. 2016. Twitter-network topic model: A full Bayesian treatment for social network and text modeling. *arXiv preprint arXiv:1609.06791* (2016).

[29] Kuan-Yu Lin and Hsi-Peng Lu. 2011. Why people use social networking sites: An empirical study integrating network externalities and motivation theory. *Computers in human behavior* (2011).

[30] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In *Mining text data*.

[31] Caroline Lo, Dan Frankowski, and Jure Leskovec. 2016. Understanding behaviors that lead to purchasing: A case study of pinterest. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[32] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and Role Discovery in Social Networks.. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

[33] Ajay Modi, Zhibo Sun, Anupam Panwar, Tejas Khairnar, Ziming Zhao, Adam Doupé, Gail-Joon Ahn, and Paul Black. 2016. Towards automated threat intelligence fusion. In *Proceedings of the IEEE International Conference on Collaboration and Internet Computing (CIC)*.

[34] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. 2010. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*.

[35] Marti Motoyama, Damon McCoy, Kirill Levchenko, Stefan Savage, and Geoffrey M Voelker. 2011. An analysis of underground forums. In *Proceedings of the ACM SIGCOMM Conference on Internet Measurement (IMC)*.

[36] Mor Naaman, Jeffrey Boase, and Chih-Hui Lai. 2010. Is it really about me?: message content in social awareness streams. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW)*.

[37] Minh-Thap Nguyen and Ee-Peng Lim. 2014. On predicting religion labels in microblogging networks. In *Proceedings of the ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR)*.

[38] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, Noah A Smith, et al. 2010. From tweets to polls: Linking text sentiment to public opinion time series.. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

[39] Jaziar Radianti. 2010. A study of a social behavior inside the online black markets. In *Proceedings of the International Conference on Emerging Security Information, Systems and Technologies (SECURWARE)*.

[40] Craig Ross, Emily S Orr, Mia Sisic, Jaime M Arseneault, Mary G Simmering, and R Robert Orr. 2009. Personality and motivations associated with Facebook use. *Computers in human behavior* (2009).

[41] Salvatore Scellato, Anastasios Noulas, and Cecilia Mascolo. 2011. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[42] Gwendolyn Seidman. 2013. Self-presentation and belonging on Facebook: How personality influences social media use and motivations. *Personality and Individual Differences* (2013).

[43] Brett Stone-Gross, Thorsten Holz, Gianluca Stringhini, and Giovanni Vigna. 2011. The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns. *LEET* (2011).

[44] 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 *Proceedings of the USENIX Security Symposium (USENIX)*.

[45] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.

[46] Ziming Zhao, Gail-Joon Ahn, Hongxin Hu, and Deepinder Mahi. 2012. SocialImpact: systematic analysis of underground social dynamics. In *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*.

[47] Ziming Zhao, Mukund Sankaran, Gail-Joon Ahn, Thomas J Holt, Yiming Jing, and Hongxin Hu. 2016. Mules, Seals, and Attacking Tools: Analyzing 12 Online Marketplaces. *IEEE Security & Privacy* (2016).

[48] Juan Zheng, Zhimin He, and Zhe Lin. 2017. Hybrid adversarial sample crafting for black-box evasion attack. In *Proceedings of the IEEE International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*.

[49] Yilu Zhou, Edna Reid, Jialun Qin, Hsinchun Chen, and Guanpi Lai. 2005. US domestic extremist groups on the Web: link and content analysis. *IEEE intelligent systems* (2005).