A Heuristic Method for Identifying Scam Ads on Craigslist

Hamad Alsaleh
Department of Information Systems
University of Maryland, Baltimore County
Baltimore,MD, USA
Alsaleh1@umbc.edu

Lina Zhou
Department of Information Systems and Operations Management
The University of North Carolina at Charlotte
Charlotte, NC, USA
lzhou8@uncc.edu

Abstract- **Craigslist is a popular online customer-to-customer marketplace, which has attracted millions of consumers for trading and purchasing secondhand items. Because of the high financial return that sellers could gain from using this site and the anonymity option that the website provides to its users, Craigslist is highly subject to fraudulent activities. The primary objective of this study is to detect scam ads on Craigslist. Based on the related literature and our observations of ads collected from the platform, we develop a heuristic method for identifying scam ads. We evaluate the proposed heuristics by conducting an experiment and performing additional data analyses using real data. The results provide preliminary evidence for efficacy of the heuristics developed in this study.**

Keywords – **Scam, heuristics, online ads, flagged ads, Craigslist**

I. INTRODUCTION

Craigslist is a site of online classified ads that serves as an alternative to the traditional medium for selling products, which minimizes the effort and reduces the time required to reach other buyers. The site also offers a wider reach, easier access, and a cheaper method of advertising in comparison with traditional media [1]. These characteristics of Craigslist makes it appealing to potential online fraud such as scam ads.

Scams are fraudulent schemes that could range from stealing money to a possibility of being life threatening. Craigslist offers tips for how to avoid scams and even provide examples of various types of scam activities. Additionally, it flags ads based on users voting by implenting a community moderation detection system where visitors or potential buyers play an active role. However, there have been widespread reports of scam incidents from users who fell victims to Craigslist scams [2, 3].

Research studies on Craigslist scams remain scarce. Some of them have solely relied on flagged (banned) ads to categorize scam ads[4,5]. However, these limited studies overlooked the differences between scam and spam ads.

Flagged ads are banned because they: 1) sell prohibited items such as weapons, drug and other prohibited items listed on the site [6], 2) are spam ads, or 3) are posted by authors who are suspected of being scammers. Thus, all flagged ads are not scams. and vice versa. This research aims to address the above limitations by developing a heuristic method for identifying scam ads on Craigslist. In this study, we will also assess the effectiveness of Craigslist in banning spam ads, The findings of this study shed light on the nature of flagged ads and highlight the need of deepening our understanding of scam ads to improve the state of practice in scam detection. The heuristics will pave the way for building computer models for the support of scam detection on Craigslist.

The rest of the paper is organized as the following. In the next section, we provide background on Craigslist ads and review related work. Then, we provide key concepts related to this study, and develop heuristics for identifying scam ads from the platform. Subsequently, we provide descriptive statistics of the selected data. Finally, we discuss the findings of the study and their implications.

II. BACKGROUND AND RELATED WORK

The majority of ads posted on Craigslist are free, and the life span of posts on the site vary depending on other factors such as the section that an ad is posted in. In the for-sale section, an ad post could stay active for a maximum of 30 days in metropolitan areas such as Boston and Washington D.C. After that, the ad would automatically became expired. The difference in the active period is mainly attributed to the amount of traffic to the different local sites.

Scams on Craigslist occur most frequently in the vehicles, rental, ticket, and personnel sections, as well as on products with high value due to the latter's financial motivations [7]. Since posting is free in most sections on the site, scammers are motivated to post multiple scam ads without any additional cost.

There is a paucity of literature on combating the issue of fraud or specifically scam on Craigslist to date. Tran et al. [4] presented a method for detecting spam ads in the car section of the Craigslist. In their work, they identified spam posts as advertisements posted multiple times. In addition, they used human evaluators to label a sample of 500 spam posts. Park et al. [8] conducted a study to identify rental scam campaigns on the site. Their approach was to first characterize rental scam from a survey of online reports of rental scammers. They built their heuristics based on the correlation of listings posted on Craigslist with other rental sites and a noticeable price difference. Nilizadeh et al. [9] introduced various heuristics to identify scam campaigns across metropolitan cities in the U.S. in the car and truck sections on Craigslist. Garg and Niliadeh [5] examined the relationship between automobile scam on Craigslist and cultural, structural, and economical characteristics of communities in the U.S. However, these studies fail to either differentiate scam from spam ads or differentiate scam from flagged ads. This study will address the above limitations by developing a heuristic method for identifying scam ads on Craigslist.

III.   THE DEVELOPMENT OF HEURISTICS FOR IDENTIFYING SCAM ADS

Scam is defined as a fraudulent scheme where an individual or group of individuals misleads potential buyers by offering them products that don't exist or by presenting false information about the product's true condition. The goal of scamming is to gain money without sending the product to the buyer [10].

A.   Data Collection

We chose the cellphone section of Craigslist to examine scam ads. Prior to data collection, we explored the length of time required to collect sufficient ads based on the volume of traffic. We collected data from a list of metropolitan cities in various regions of the U.S. such as Boston, Atlanta, San Francisco, Baltimore, and Detroit. The sites of these metropolitan cities tend to attract high volumes of ads based on online reports [11] and our observations. Therefore, selecting these cities will increase the chance of collecting scam ads.

The dataset contained a total of 81,505 ads. Based on their status, these ads were categorized into four types: active, flagged, removed, and expired. Active ads remained active and could be viewed by any visitor to the site. For flagged ads, they contain the following message, "This posting has been flagged for removal". As explained earlier, Craigslist relied on its users to detect suspicious ads by voting on such ads. Removed ads have been removed by their posters. Lastly, expired ads were hidden by the website, which were

labelled with the following message: "This posting has expired".

Among these ads, 5827 (7.1%) were flagged, 18732 (23%) were active, 25945 (31.8%) were removed, and the remaining were expired ads. The statuses of ads were based on their statuses at the time of terminating our data collection.

B.   Data Cleaning and Preprocessing

Data cleansing is an important precursor to an in-depth analysis of online ads. We addressed data noise issues, including data errors, different formats, and missing values related to phone specs such as model, brand, and phone condition. We developed strategies for addressing each type of the data noise.

The distribution of ads after data cleaning and preprocessing is reported in Fig.1. The figure shows that about 4.3% ads in the cleaned dataset were flagged. Nevertheless, these ads were insufficient for detecting scam. For instance, they likely contained a mixture of scam ads and trustworthy ads, and the latter may be flagged with the underlying motivation of competition. Additionally, they may also miss other scam ads that were not flagged. Therefore, we developed heuristics for identifying scam ads.
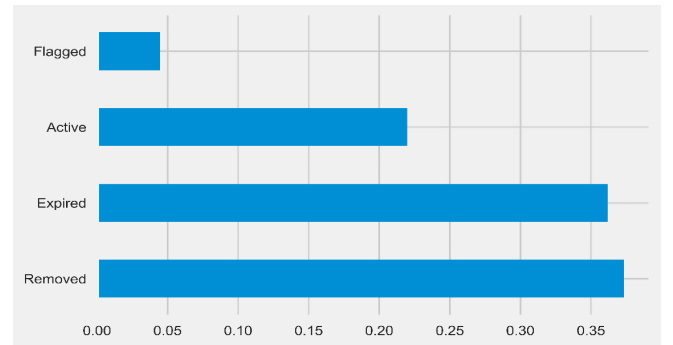

Fig.1. Percentage distribution of ads across different status

C.   Heuristics for Determining Scam Ads

To identify scam ads, we built heuristics based on findings of previous studies [8, 9], examples provided by Craigslist [12], and our own observations of the real data. Specifically, we started with flagged ads. This is because flagged ads have a higher tendency to be scam than other types of ads [9]. In addition, scam ads should be distinct from trustworthy ads. Thus, there is a need to develop method for differentiating scam ads from trustworthy ads.

We chose price as one of the main factors that signal scam ads. The heuristic was developed based on the following rationales: 1) it is unusual for someone to advertise products with significantly high or low price; 2) the price of a product is referenced as a rule to identify scam ads on Craigslist [12]; and 3) several studies used price as a factor in identifying scams for other product categories [8,9].

There are practical challenges with implementing the price heuristic. To this end, we designed a multi-step process to set the threshold for price. We first clustered cellphones by model. Then, we analyzed the price distributions of phone models within each cluster and calculated the median price for each cluster. We further examined the percentile distribution of price across the ads of different status categories.

As shown in Fig.2, the flagged ads have a relatively higher concentration in the two extreme ranges (i.e., between 0 to 10th and between 90 to 100th percentiles) than in the middle ranges. In addition, the removed ads exhibit an opposite pattern. Therefore, scam ads were more prevalent with extremely cheap and highly priced phone models. The data descriptive statistics provides strong evidence for the validity of using price as heuristic for establishing the ground truth for scam ads.
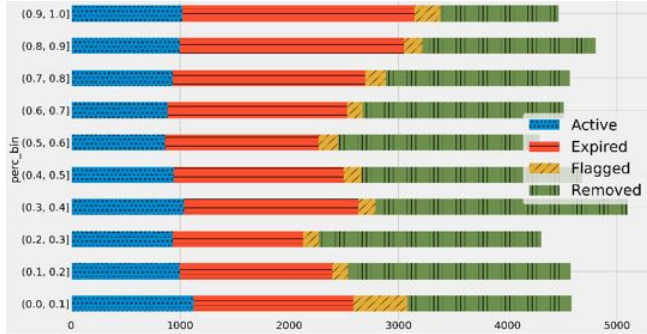

Fig.2. Price Distribution across the Ads of Different Status Categories

We also developed heuristics to identify additional scam ads. For instance, if a suspicious ad was highly similar to an identified scam ad, the former ad was treated as scam. In other words, using the identified scam ads as seeds, we further grew the seed set by including other similar ads. In the similarity measurement, we assigned higher weight to some parts of an ad such as phone number than other parts.

We applied the proposed heuristics to create a scam dataset through the following procedure (see Fig.3)

- select all the ads that fall below 10th percentile and above 98th percentile of price as candidate scam ads except for those in 'removed' status;
- select all the ads with 'flagged' status as scam ad and label the rest of ads in the candidate set as 'suspicious'.
- Filter the set of suspicious ads using the similarity heuristic.

As a result, we generated a dataset of 2,091 scam ads. Among them, only 493 were flagged by Craigslist and the remaining were not.
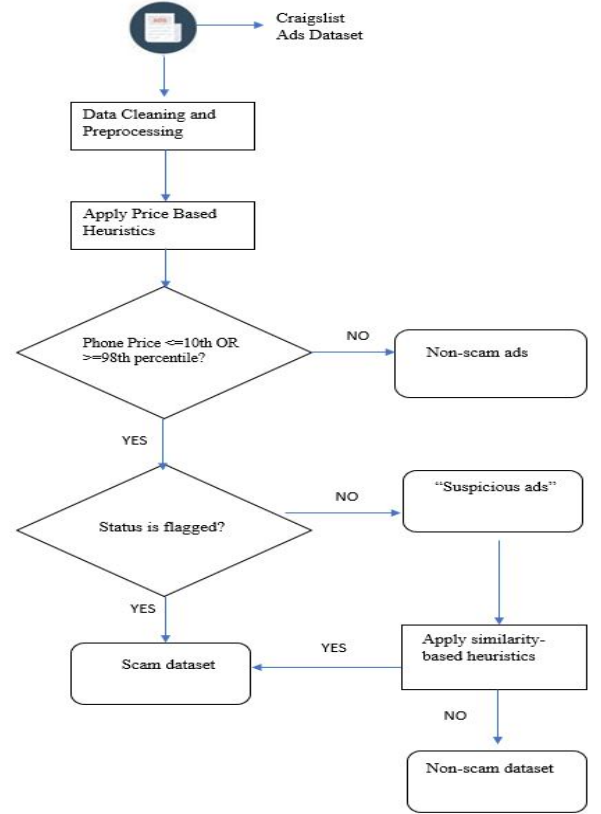

Fig.3. A heuristic method for detecting Scam Ads

IV. A PRELIMINARY EVALUATION

To evaluate the proposed heuristics, we conducted a small-scaled field experiment. We contacted 30 suspects from our scam dataset by email. In the email, we inquired about the status of the phones and the preferred method of payment. We received 10 email responses.

We performed an in-depth analysis of the responses, and identified some interesting patterns.

- The majority of the responses were sent from email addresses that were different from the ones we originally contacted.

- All the responses referred to different URLs where we could get these phones for a significantly lower price. Our checking of these sites revealed that they were penny auction websites with a new domain name.

- The responses shared similar content and writing style.

These observations provide preliminary evidence for validity of the identified scam ads.

To further and directly verify the identified heuristics, particularly related to the price, we conducted another experiment. In this experiment, we compared the asking price in the identified scam ads against the corresponding market price. The market price was derived based on the specs of the phones such as brand, model and year. To quantify the price deviation from the market price, we developed a metric called price ratio, which was defined as the ratio of the asking price to the market price of the advertised phone.

Our analyses of the selected scam ads in terms of price ratio showed that over 70% of scam ads asked for prices less than 57% of the market price of the advertised phones. In addition, we found that around 80% of the selected scam ads asking for significantly high prices had the price ratio above 150%. These results confirm the utility of the price heuristic in detecting scam ads.

V. DISCUSSION

In this study, we developed heuristics for the identification of scam ads in Craigslist based on price and similarities between the scam ads and other suspicious ads. By applying the heuristics, we analyzed a large set of online ads collected from Craigslist. The evaluation results based on email contact and price comparison provided preliminary evidence for the efficacy of the heuristics for identifying scam ads.

Our heuristic for identifying scam ads is consistent with Craigslist's guidelines for avoiding scams [12]. The analysis of flagged ads in different price ranges reveals high concentrations of flagged ads in the lowest and highest percentiles. More importantly, our findings reveal that some of the scam ads were not flagged on the platform, which highlight the importance of improving scam detection.

This study makes several research contributions and has practical implications. First, it introduces a first set of heuristics for the identification of scam ads in the cellphone section of Craigslist. Second, given the very limited studies on Craigslist ads, there is a lack of information about the structure and management of ads on the platform. This study lays the foundation for future research by offering an in-depth examination of data collected from on the platform. Third, the findings of this study can provide concrete suggestions for potential buyers on Craigslist to avoid scam. Among others, interested buyers should check the price of advertised products before making a purchase decision. The heuristics developed in this study can be extended to detect scam ads for other types of products that share similar characteristics to cellphones on Craigslist.

This work can be continued in several ways. One possible extension is to conduct additional validation by contacting more suspected scam sellers and even characterize scam campaigns. Another possible extension of this work is to develop computational models for the automatic detection of online scams. The proposed heuristics and models can be implemented as a browser plugin to provide decision support for the detection of scam ads. Finally, it is promising to conduct content analyses of scam ads to gain deeper insights into their nature.

REFERENCES

[1] Roennevig, M. (n.d.). What Are the Advantages of Advertising in Classifieds?.*chron.com*.[Online].Available: www.smallbusiness.chron.com/advantages-advertising-classifieds 57375.html.[Accessed:September -12-2018].

[2] Muschick, P.New twist on Craigslist rental scam involves costly credit reports.*mcall.com*.2017.[Online].Available:https://www.mcall.com/news/local/watchdog/mc-craigslist-rental-scam-credit-score-watchdog-20170201-column.html.[Accessed:September-12-2018].

[3] Kearns , S. (n.d.). 7 Types of Common Cragislist Scams to Watch Out for.*moneycrashers.com*.[Online].Avalaible.: www.moneycrashers.com/types-common-craigslist-scams/ [Accessed: September-12-2018].

[4] Tran, H., Hornbeck, T., Ha-Thuc, V., Cremer, J., & Srinivasan, P. (2011, March). Spam detection in online classified advertisements. In *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality* (pp. 35-41). ACM.

[5] Garg, V., & Niliadeh, S. (2013, May). Craigslist scams and community composition: Investigating online fraud victimization. IEEE Symposium on Security and Privacy Workshops (SPW),San Fransisco,CA, 2013 IEEE (pp. 123-126).

[6] Prohibited.(n.d.).*Craigslist.com*.[Online].Available:https://www.craigslist.org/about/prohibited.[Accesseed: September-12-2018]

[7] Bonser, C. (n.d.). How to View Someone's Profile on Craigslist., *techwalla.com*.[Online].Available: www.techwalla.com/articles/how-to-view-someones-profile-on-craigslist.[Accessed:September-12 2018].

[8] Park, Y., McCoy, D., & Shi, E. (2016). Understanding Craigslist Rental Scams. In International Conference on Financial Cryptography and Data Security (pp. 3-21). Springer, Berlin, Heidelberg.

[9] Nilizadeh, S., Orlova, D. A., Nematzadeh, A., Gupta, M., Kapadia, A. C., & Kalafut, A. In Search of a Toyota Sienna: Prevalence and Identification of Auto Scam on Craigslist.

[10] McCormick, A. M., & Eberle, W. (2013, May). Discovering Fraud in Online Classified Ads. In Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, St. Pete Beach, FL.

[11] B. Craigslist Most Trafficked Cities .*craigslistpostingservice.net* .2017.[Online].Available:https://www.craigslistpostingservice.net/craigslist-most-trafficked-cities-2017-edition/[Accessed:September-12-2018].

[12] scams.(n.d.).*Craigslist.com*.[Online].Available: www.craigslist.org/about/scams.[Accessed: September-12-2018].