2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Detecting Malicious Campaigns in Crowdsourcing Platforms

Hongkyu Choi
Department of Computer Scinece
Utah State University
Logan, UT 84322
hongkyu.choi@aggiemail.usu.edu

Kyumin Lee
Department of Computer Science
Utah State University
Logan, UT 84322
kyumin.lee@usu.edu

Steve Webb
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
steve.webb@gmail.com

***Abstract*—Crowdsourcing systems enable new opportunities for requesters with limited funds to accomplish various tasks using human computation. However, the power of human computation is abused by malicious requesters who create malicious campaigns to manipulate information in web systems such as social networking sites, online review sites, and search engines. To mitigate the impact and reach of these malicious campaigns to targeted sites, we propose and evaluate a machine learning based classification approach for detecting malicious campaigns in crowdsourcing platforms as a first line of defense. Specifically, we (i) conduct a comprehensive analysis to understand the characteristics of malicious campaigns and legitimate campaigns in crowdsourcing platforms, (ii) propose various features to distinguish between malicious campaigns and legitimate campaigns, and (iii) evaluate a classification approach against baselines. Our experimental results show that our proposed approaches effectively detect malicious campaigns with low false negative and false positive rates.**

I. Introduction

Crowdsourcing platforms such as Mechanical Turk (MTurk) and Crowdflower provide a marketplace where requesters recruit workers and request the completion of various tasks. Since anyone in the world can be a worker, the labor fees are relatively low, and workers are available at virtually all hours of the day. Due to these benefits, requesters have used crowdsourcing platforms for various tasks such as labeling datasets, searching a boat from satellite images to find a lost person, proofreading a document, and adding missing data.

However, some requesters abuse crowdsourcing platforms by creating malicious campaigns to manipulate search engines, write fake reviews, and create accounts for additional attacks. Using crowdsourced manipulation, malicious requesters and workers can potentially earn hundreds of millions of dollars. As a result, crowdsourced manipulation threatens the foundation of the free and open web ecosystem by reducing the quality of online social media, degrading trust in search engines, manipulating political opinion, and ultimately compromising the security and trustworthiness of cyberspace [1]–[3].

Prior research [2], [3] identified the threat of malicious campaigns by quantifying their prevalence in several crowdsourcing platforms. Specifically, a large collection of loosely-moderated crowdsourcing platforms serves as launching pads for these malicious campaigns. Unfortunately, there is a significant gap in our understanding of how to detect malicious campaigns at the source (i.e., crowdsourcing platforms), which would mitigate their impact and reach before they influence targeted sites.

Hence, in this paper we aim to automatically predict and detect malicious campaigns in crowdsourcing platforms by answering following research questions: What kind of malicious campaigns exist in crowdsourcing platforms? Can we find distinguishing patterns/features between malicious campaigns and legitimate campaigns? Can we develop a statistical model that automatically detects malicious campaigns?

To answer these questions, we make the following contributions:

- First, we collect a large number of campaigns from popular crowdsourcing platforms: MTurk, Microworkers, Rapidworkers, and Shorttask[1]. Then, we cluster malicious campaigns to understand what types of malicious campaigns exist in crowdsourcing platforms.
- Second, we analyze characteristics of malicious campaigns and legitimate campaigns in terms of their market sizes and hourly wages. Then, we propose and evaluate various features for distinguishing between malicious campaigns and legitimate campaigns, and we visualize each feature to concretely illustrate the differing properties for malicious and legitimate campaigns.
- Third, we develop a predictive model, and evaluate its performance against baselines in terms of accuracy, false positive rate and false negative rate. To our knowledge, this is the first study to focus on detecting malicious campaigns in multiple crowdsourcing platforms.

II. Related Work

Since the emergence of crowdsourcing platforms (e.g., MTurk and Crowdflower), researchers have studied how to use crowd wisdom. Wang et al. [4] hired workers to identify fake accounts in Facebook and Renren. Workers have identified

[1]MTurk, Microworkers, Rapidworkers and Shorttask represent www.mturk.com, microworkers.com, rapidworkers.com, and shorttask.com, respectively.

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA
978-1-5090-2846-7/16/$31.00

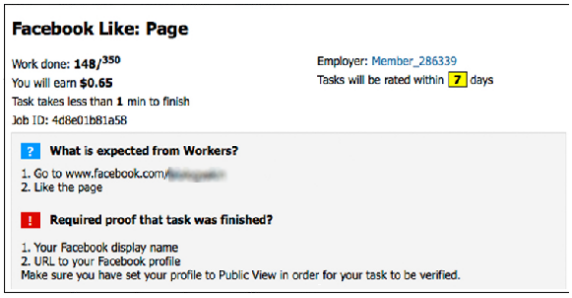


Fig. 1. A campaign description.

improper tasks in a Japanese crowdsourcing site [5] and proofread documents in near-real time [6]. Other researchers were interested in analyzing the demographics of workers [7] and quantifying the evolution of campaigns/tasks in MTurk [8]. Ge et al. analyzed a supply-driven crowdsourcing marketplace regarding key features that distinguish "super sellers" from regular participants [9].

Another research topic is to measure the quality of workers and outcomes (and determining how to control that quality). Venetis and Garcia-Molina [10] proposed three scoring methods such as gold standard, plurality answer agreement, and Task Work Time to filter low quality answers. A machine learning technique was applied to detect low quality answerers [11]. Soberón et al. [12] showed that adding open-ended questions (i.e., explanation-based techniques) into tasks was useful for identifying low quality answers.

With the rising popularity of crowdsourcing systems, malicious campaigns and tasks have been created by some requesters. To understand the problems, Motoyama et al. [2] introduced possible web service abuse in Freelancer.com. Wang et al. [3] analyzed two Chinese crowdsourcing platforms and found that up to 90% of campaigns are malicious campaigns. Lee et al. [1] found that social networking sites and search engines were mainly targeted by malicious campaigns. Researchers began analyzing crowdsourced manipulation and the characteristics of workers in targeted sites such as Facebook and Twitter. Fayazi et al. [13] proposed a reviewer-reviewer graph clustering approach based on a Markov Random Field to identify workers that posted fake reviews on Amazon.

In contrast to this previous research, we collected a large number of campaigns from four crowdsourcing platforms, analyzed characteristics of malicious campaigns and legitimate campaigns, and developed predictive models to automatically identify malicious campaigns. Our research will complement existing research base.

III. Dataset

In a crowdsourcing platform, there are two types of users – (i) a requester and (ii) a worker. A requester is a user who creates a campaign with detailed instructions for one or more tasks. Each task is then performed by one worker. If the requester is satisfied with the worker's outcome, the requester will approve it, and compensation (i.e., money) will be passed to the worker by the crowdsourcing platform.

TABLE I
Statistics for tasks and market sizes of malicious campaigns and legitimate campaigns.

Malic. T.	Legit T.	Malic. M.	Legit M.
798,796	2,557,357	$148,911	$179,696

To collect a dataset, we developed a crawler for four popular crowdsourcing platforms: Amazon Mechanical Turk (MTurk), Microworkers, Rapidworkers, and Shorttask. The crawler collected campaign listings and detailed campaign descriptions. We ran the crawler for 3 months between November 2014 and January 2015, and it collected 23,220 campaigns consisting of 3,356,153 tasks[2]. Figure 1 shows a sample campaign description which contains the number of available tasks, compensation for each task, estimated time to complete a task, and task instructions that describe what a worker is supposed to do.

We define a malicious campaign as one that requires workers to manipulate information in targeted sites such as social media sites and search engines. For example, a malicious campaign might require workers to post fake reviews on Amazon, artificially create backlinks to boost a specific website's search engine ranking, or "Like" a specific Facebook page.

Using this definition, two annotators manually labeled each campaign in our dataset as a malicious campaign or a legitimate campaign, based on the campaign description. When the two annotators disagreed about a particular campaign's label, a third annotator labeled the campaign. The annotators made the same labeling decision on 23,079 out of 23,220 campaigns, achieving 99.4% agreement.

Our collected dataset consisted of 5,010 malicious campaigns and 18,210 legitimate campaigns. Each campaign contained 145 tasks on average. Overall, the malicious campaigns contained about 800K tasks.

IV. Characteristics of Malicious Campaigns and Legitimate Campaigns

Now we turn our attention to analyzing characteristics of malicious campaigns and legitimate campaigns in the crowdsourcing platforms.

Tasks and market sizes. First, we calculated the number of tasks associated with malicious campaigns and legitimate campaigns, and then, we measured the market sizes for malicious and legitimate campaigns. To estimate the market size for a collection of malicious and legitimate campaigns, we used the following equation:

$$MarketSize(C) = \sum_{i=1}^{n} r(i) * count(i) \quad (1)$$

, where C is a set of malicious or legitimate campaigns $\{c_1, c_2, c_3 ... c_n\}$ in crowdsourcing platforms, n is the number of malicious or legitimate campaigns, $r(i)$ is the reward (i.e.,

[2]A campaign contains multiple tasks, and each task is assigned to one worker.

TABLE II
GOALS AND MEDIAN VALUES OF MALICIOUS CAMPAIGN CLUSTERS.

Campaign Goal	Malic. C.	%	ETC (min)	Reward	$ / hour
Social network associated (Review, Link, Share, Retweet and Like)	2,987	60	33.5	$0.45	$2.13
Search and click	863	17	8	$0.22	$2.65
Search and visit	654	13	5	$0.21	$3.85
Add a comment	197	4	8	$0.31	$2.91
Register in a forum and post a message	168	3	12	$0.35	$1.02
Create a new pin at Pinterest	96	2	3	$0.11	$2.20
Download and install a new application	45	1	12.5	$0.81	$4.50
Average			12	$0.35	$2.75

compensation) per task for campaign i, and $count(i)$ is the number of tasks associated with campaign i.

As shown in Table I, the malicious tasks for the four crowdsourcing platforms amounted to 45% of the entire market size, while the number of malicious tasks only represented 24% of the total campaign tasks. This analytical result reveals that the reward per malicious task is much higher than the reward per legitimate task.

Hourly wages. Next, we concretely evaluate if the hourly reward for malicious campaigns is actually higher than the hourly reward for legitimate campaigns. A campaign's description contains estimated time to complete (ETC) information and a reward per task. We calculated the hourly wage for each campaign using $\frac{reward\,per\,task*60}{ETC}$ because ETC's unit of measurement is a minute. The median hourly wage in malicious campaigns ($2.48) is larger than the median hourly wage in legitimate campaigns ($1.88). One explanation for this result is that malicious requesters provide higher rewards to workers so that they can attract these workers, who may have ethical concerns about these malicious campaings/tasks.

Clustering malicious campaigns. To investigate characteristics of malicious campaigns associated with specific goals, we clustered the campaigns based on their goals and targeted sites. From 5,010 malicious campaigns, we extracted a title from each campaign and tokenized it by unigram. Then, we removed stop words and measured TF-IDF. Now, each campaign is represented as a vector based on TF-IDF. Given a list of vectors, we used a k-means clustering algorithm to cluster the vectors (i.e., campaigns). To obtain the optimal number of clusters, we experimented with k values in the range of 2 through 10, and we measured Sum of Squared Error (SSE) for each value. We found that $k = 7$ was the optimal cluster size.

After clustering the malicious campaigns, we investigated every cluster and found objectives for the campaigns in each cluster as shown in Table II. The median values for time, reward, and hourly wage are presented in the table. The goals for most of the campaigns were to manipulate content on social networking sites (e.g., Google Plus, Twitter, Yahoo Answer and Facebook) and manipulate search engine results by searching a specific keyword and clicking a certain web page link. "Download and install a new application" campaigns provided workers with larger compensation per hour than the other campaigns.

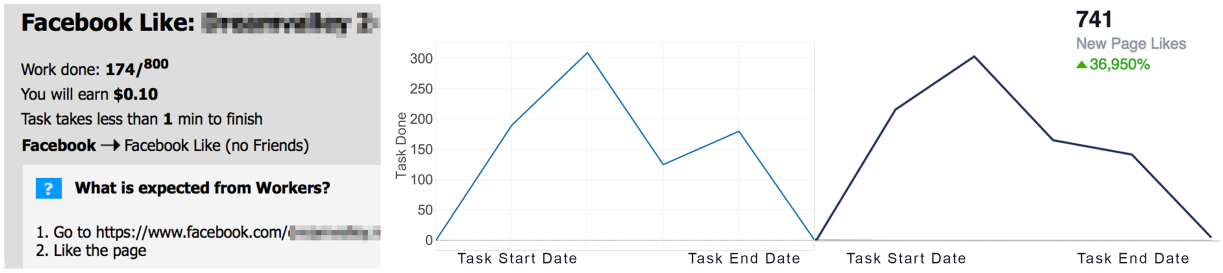
TABLE III
THE FIVE MOST TARGETED SITES BY MALICIOUS CAMPAIGNS AND THEIR CORRESPONDING MEDIAN VALUES.

	Malic. C.	%	Reward	ETC(min)	$/hour
Google	902	18	$0.21	6.0	2.3
Twitter	600	12	$0.16	7.5	1.9
Instagram	210	4	$0.13	6.5	1.0
Facebook	154	3	$0.35	7.0	3.0
Youtube	153	3	$0.20	9.5	2.2

To identify the sites that were targeted the most by malicious campaigns, we extracted a list of the top 500 companies from Alexa, and we searched for each of those company names (and company hostnames) in malicious campaign descriptions. Table III shows the five most targeted sites. 902 (18%) malicious campaigns targeted Google. Social networking sites such as Twitter, Instagram, Facebook, and Youtube were also targeted frequently.

Real-world impact of malicious campaigns. Thus far, we've identified important characteristics of malicious campaigns. Now, we need to determine if malicious campaigns have any real-world impact on targeted sites and if existing security algorithms/systems can detect manipulations in the targeted sites. To investigate these issues, we tracked 29 malicious campaigns targeting Facebook in which workers manipulated Facebook likes. We collected daily snapshots of the malicious campaigns from crowdsourcing platforms and daily snapshots of the targeted Facebook pages. The 29 malicious campaigns consisted of 8,268 tasks, each task required adding one fake Like. Out of 8,268 fake likes, 7,160 of the likes were successfully attributed to the target pages when we checked those pages later, which means only 1,108 (13.4%) of the fake likes were deleted by the Facebook security team.

Figure 2 shows an example of the malicious campaigns manipulating the number of Facebook likes. The left figure shows the campaign description containing a total number of tasks, the number of available tasks, and task instructions for workers. 800 of the campaign's tasks were completed within 4 days. The middle figure and right figure show the number of completed tasks reported to a requester and the number of fake likes completed by workers for the targeted Facebook page. We can clearly observe that the middle and right figures show similar temporal patterns. Out of 800 likes, 741 likes



Actual Task Description targeting on Facebook like | Trend of Task Completion after Task Posting | Changes in the number of likes in Facebook page

Fig. 2. Consequence of manipulating the number of likes in Facebook.

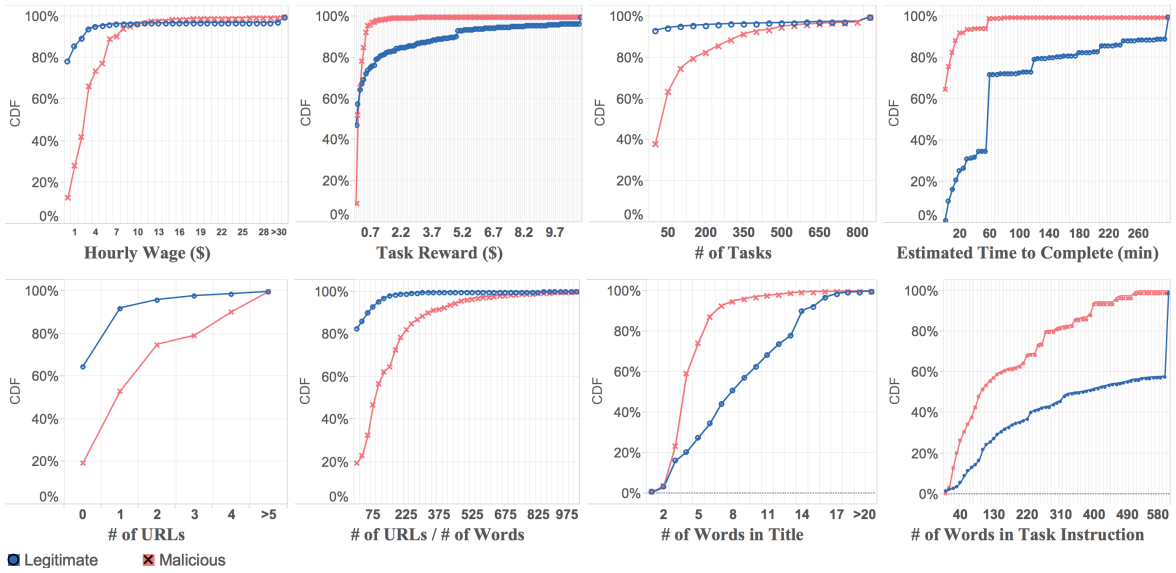


Fig. 3. Cumulative distribution function of features by class (○: legitimate campaigns, ×: malicious campaigns).

remained on the Facebook page, which means Facebook only labeled 59 (7%) likes as "fake" likes.

This example and the previous analysis (for 29 fake liking campaigns) show that malicious campaigns have a real-world impact on targeted sites, and current security systems are unable to detect most of the manipulated content. The previous work [14] also confirmed that Twitter safety team only detected 24.6% of fake followers. These results motivated us to investigate an automated approach for detecting malicious campaigns in crowdsourcing platforms using predictive models.

V. Features

In this section, we describe proposed features for building malicious campaign classifiers. To build a universal classifier which can be applied to any crowdsourcing platform regardless what information is available, we proposed and extracted commonly available features across the four crowdsourcing platforms. Our proposed features are reward, number of tasks, estimated time to complete (ETC), hourly wage, number of URLs in task instruction, $\frac{Number\ of\ URLs\ in\ task\ instruction}{Number\ of\ words\ in\ task\ instruction}$, number of words in a task title, number of words in task instruction, and text features extracted from task title and task instruction.

To avoid the overfitting problem by removing features that are too similar, we measured the Pearson correlation coefficient of each pair of the first 8 features excluding text features. We kept the 8 features because there was no significant correlation.

From task title and task instruction, we extracted text features as follows: (i) first, we removed stopwords from the title and task instruction, and then, we applied stemming to them; (ii) second, we extracted unigrams, bigrams, and trigrams from the text; (iii) third, we measured χ^2 values for the extracted unigram, bigram, and trigram features; (iv) finally, we only used text features with positive χ^2 values. Through this process, we used thousands of text features.

Next, Figure 3 shows cumulative distribution functions (CDFs) for malicious campaigns and legitimate campaigns. Interestingly, requesters for 80% of the legitimate campaigns paid less than one dollar to each worker in terms of hourly wage, while requesters for 10% of the malicious campaigns paid the same hourly wage to workers. This suggests that performing malicious campaigns was more profitable, which is consistent with our previous results.

Malicious campaigns also contain a larger number of tasks than most legitimate campaigns, and malicious campaigns have shorter ETC than legitimate campaigns. Task instructions in malicious campaigns contain more URLs than legitimate campaigns, which suggests that malicious campaigns require workers to access external websites (potentially targeted sites) more often.

Finally, malicious campaigns have shorter titles and task instructions than legitimate campaigns. This observation might indicate that some of the legitimate campaigns are more complicated to perform and require longer ETC. Overall, the CDFs illustrate distinct differences between malicious campaigns and legitimate campaigns.

VI. Experiments

In the previous section, we observed that malicious campaigns and legitimate campaigns have different characteristics. In this section, we build classifiers to detect malicious campaigns by exploiting these differences.

A. Detecting Malicious Campaigns

As we mentioned in Section III, we collected campaign descriptions for 3 months between November 2014 and January 2015. The dataset consists of 18,210 legitimate campaigns and 5,010 malicious campaigns. We built and tested statistical models with 10-fold cross validation. We compared the performance of three classification algorithms: Naive Bayes, J48, and Support Vector Machine (SVM).

We compared our statistical models/classifiers with following baselines: (i) *majority selection approach* which always predicts a campaign's class as the majority instances' class (i.e., a legitimate campaign in the dataset); (ii) *URL-based filtering* approach which classifies a campaigns as a malicious campaign if its description contains at least one URL whose host name is one of top K sites; and (iii) *principal component analysis* (PCA) approach, an unsupervised machine learning technique, inspired from the previous work [15]. In PCA approach, we projected campaigns (using the same features with our classifiers) onto the normal and residual subspaces to classify malicious and legitimate campaigns. The space spanned by top principal components is the normal subspace and the remaining space is known as residual subspace. From our dataset, we achieved 85% variance from the top 35 principal components out of 1,835 components. We computed L2 norm and set the squared prediction error as the threshold value to find the malicious campaigns. We changed the threshold value from 1% to 50% by 1% increment each time to get the best classification result. Campaigns, whose L2 norm was greater than the threshold value, were classified as malicious campaigns.

To evaluate the performance of classifiers, we used following evaluation metrics: accuracy, false positive rate (FPR) and false negative rate (FNR). FPR means malicious campaigns were misclassified as legitimate campaigns while FNR means legitimate campaigns were misclassified as malicious campaigns.

TABLE IV
Classification results.

Approach	Accuracy	FPR	FNR
Majority Selection	78.4%	1	0
URL-based filtering@100	72.4%	0.708	0.157
URL-based filtering@500	72.3%	0.688	0.164
URL-based filtering@1000	71.9%	0.635	0.183
PCA - 12% threshold	85.2%	0.999	0.031
our Naive Bayes	89.0%	0.044	0.147
our J48	99.1%	0.023	0.058
our SVM	**99.2%**	0.019	0.055

In experiments, we ran majority selection approach, URL-based filtering approach at top 100, 500 and 1000 sites, PCA approach, and our three classification approaches (Naive Bayes, J48 and SVM). Table IV shows experimental results of the baselines and our classification approaches. Majority selection approach achieved 78.4% accuracy, 1 FPR and 0 FNR, URL-based filtering@100 achieved 72.4% accuracy, 0.708 FPR and 0.157 FNR, and PCA approach with 12% threshold (only reporting the best result) achieved 85.2% accuracy, 0.999 FPR and 0.031 FNR. Overall, our SVM-based classifier significantly outperformed the other approaches, achieving 99.2% accuracy, 0.019 FPR and 0.055 FNR, and balancing between low FPR and low FNR.

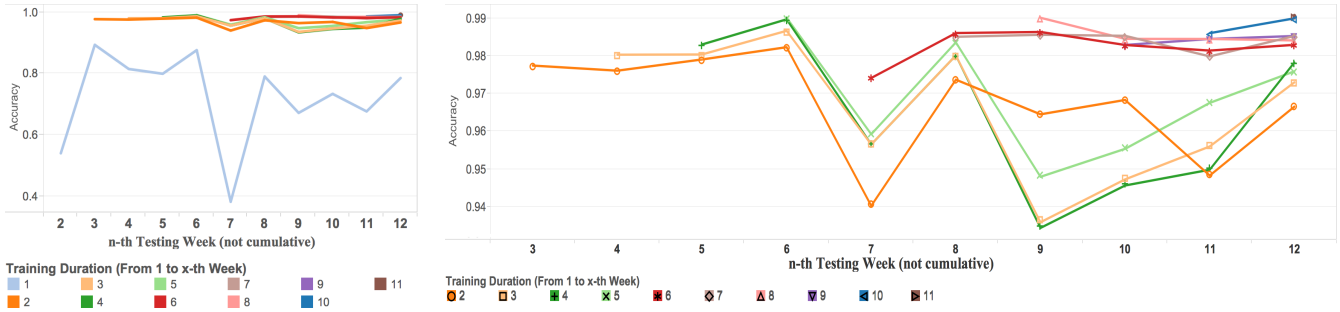
B. Robustness of Our Proposed Approach

In the previous experiment, we learned SVM classifier achieved the best prediction results for detecting malicious campaigns. Now, we analyze (i) how much training data we need to achieve a high prediction rate and (ii) whether a predictive model (i.e., a classifier) would remain robust over time.

To investigate these issues, we split the dataset chronologically based on weeks (i.e., the 3 month data was split into 12 weeks). Then, we trained a SVM classifier using the first week of data, and we used the classifier to test the data for each of the next weeks. Next, we added the following week's data (e.g., the second week of data) into the training set and tested the data for each of the next weeks. Incrementally, we added each week's data to the training set until the training set included data for the first 11 weeks.

Figure 4 shows experimental results for macro-scale and micro-scale views of our approach[3]. In particular, Figure 4(a) shows experimental results of the 2nd week to the 12th week in a macro-scale view. When we used *the first week* data as a training set and applied a classifier to each of the following weeks, the classifier achieved low accuracy. However, when we added one more week of data to a training set (i.e., the training set contained the first and second week of data), the classifier achieved significantly high accuracy. Note that 7th testing week's classification results were slightly lower than earlier testing weeks because there were very small number of legitimate campaigns posted in the 7th week (e.g., 62%

[3]We did not show FPR and FNR lines because of the limited space.



(a) A macro-scale view between the 2nd and 12th weeks.

(b) A micro-scale view between the 3rd and 12th weeks.

Fig. 4. As the training set size increases, the detection rate for malicious campaigns and legitimate campaigns also increases.

malicious campaigns and 38% legitimate campaigns in the 7th testing week vs. 11% malicious campaigns and 89% legitimate campaigns in the 6th testing week). We conjecture that the 7th week is a week containing Christmas and New Year holidays, so very less number of legitimate campaigns were created while almost same number of malicious campaigns was created compared with the 6th week.

Figure 4(b) shows experimental results in a micro-scale view by removing the first week training result (i.e., the classifier that was only trained with a single week of data). Based on this figure, we clearly observe that a SVM classifier based on data for the first 2 weeks achieved high accuracy even though the performance was up and down over time. Overall, the lowest accuracy, highest FPR and highest FNR among all the cases were 93.4%, 0.19, 0.03, respectively. Based on these experimental results, we conclude that two weeks of data is enough to build an effective predictive model for identifying malicious campaigns. We also conclude that our proposed classification approach consistently and robustly worked well over time.

VII. Conclusion

In this paper, we analyzed characteristics of malicious campaigns and legitimate campaigns. The median hourly wage in malicious campaigns ($2.48) was larger than the median hourly wage in legitimate campaigns ($1.88), tempting workers to perform malicious campaigns in targeted sites such as social networking sites, online review sites, and search engines. To measure the real-world impact of malicious campaigns, we selected Facebook Liking campaigns and found that Facebook caught only 13% fake likes. This suggests that current defense systems in targeted sites are inadequate and potentially undetected malicious campaigns are deteriorating information quality and trust.

To overcome this problem, we proposed features which were distinguished between malicious campaigns and legitimate campaigns. Then, we built malicious campaign classifiers based on the features for mitigating the impact and reach of the malicious campaigns to targeted sites. Our classifiers outperformed the baselines – majority selection, URL-based filtering and PCA approaches –, achieving 99.2% accuracy, 0.019 FPR and 0.055 FNR.

Acknowledgment

This work was supported in part by NSF grant CNS-1553035. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsors.

References

[1] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, Campaigns, and Social Media: Tracking and Revealing Crowdsourced Manipulation of Social Media," in *ICWSM*, 2013.

[2] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. M. Voelker, "Dirty jobs: The role of freelance labor in web service abuse," in *USENIX Conference on Security*, 2011.

[3] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Serf and turf: crowdturfing for fun and profit," in *WWW*, 2012.

[4] G. Wang, M. Mohanlal, C. Wilson, X. Wang, M. J. Metzger, H. Zheng, and B. Y. Zhao, "Social turing tests: Crowdsourcing sybil detection," in *NDSS*, 2013.

[5] Y. Baba, H. Kashima, K. Kinoshita, G. Yamaguchi, and Y. Akiyoshi, "Leveraging non-expert crowdsourcing workers for improper task detection in crowdsourcing marketplaces." *Expert Syst. Appl.*, vol. 41, no. 6, pp. 2678–2687, 2014.

[6] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich, "Soylent: A word processor with a crowd inside," in *UIST*, 2010.

[7] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: Shifting demographics in mechanical turk," in *CHI*, 2010.

[8] D. E. Difallah, M. Catasta, G. Demartini, P. G. Ipeirotis, and P. Cudré-Mauroux, "The dynamics of micro-task crowdsourcing: The case of amazon mturk," in *WWW*, 2015.

[9] H. Ge, J. Caverlee, and K. Lee, "Crowds, gigs, and super sellers: A measurement study of a supply-driven crowdsourcing marketplace," in *ICWSM*, 2015.

[10] P. Venetis and H. Garcia-Molina, "Quality control for comparison microtasks," in *CrowdKDD workshop*, 2012.

[11] H. Halpin and R. Blanco, "Machine-learning for spammer detection in crowd-sourcing," in *Human Computation workshop in conjunction with AAAI*, 2012.

[12] G. Soberón, L. Aroyo, C. Welty, O. Inel, H. Lin, and M. Overmeen, "Measuring crowd truth: Disagreement metrics combined with worker behavior filters," in *CrowdSem 2013 Workshop*, 2013.

[13] A. Fayazi, K. Lee, J. Caverlee, and A. Squicciarini, "Uncovering crowdsourced manipulation of online reviews," in *SIGIR*, 2015.

[14] K. Lee, S. Webb, and H. Ge, "The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing," in *ICWSM*, 2014.

[15] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove, "Towards detecting anomalous user behavior in online social networks," in *USENIX Security*, 2014.