Weaponized Crowdsourcing: An Emerging Threat and Potential Countermeasures

James Caverlee and Kyumin Lee

1 Introduction

The crowdsourcing movement has spawned a host of successful efforts that organize large numbers of globally-distributed participants to tackle a range of tasks, including crisis mapping (e.g., Ushahidi), translation (e.g., Duolingo), and protein folding (e.g., Foldit). Alongside these specialized systems, we have seen the rise of general-purpose crowdsourcing marketplaces like Amazon Mechanical Turk and Crowdflower that aim to connect task requesters with task workers, toward creating new crowdsourcing systems that can intelligently organize large numbers of people. However, these positive opportunities have a sinister counterpart: what we dub "Weaponized Crowdsourcing". Already we have seen the first glimmers of this ominous new trend—including large-scale "crowdturfing", wherein masses of cheaply paid shills can be organized to spread malicious URLs in social media (Grier, Thomas, Paxson, & Zhang, 2010; Lee & Kim, 2012), form artificial grass-roots campaigns ("astroturf") (Gao et al., 2010; Lee, Caverlee, Cheng, & Sui, 2013), spread rumor and misinformation (Castillo, Mendoza, & Poblete, 2011; Gupta, Lamba, Kumaraguru, & Joshi, 2013), and manipulate search engines. A recent study finds that 90 % of tasks on many crowdsourcing platforms are for crowdturfing (Wang et al., 2012), and our initial research (Lee, Tamilarasan, & Caverlee, 2013) shows that most malicious tasks in crowdsourcing systems target

J. Caverlee (✉)
Department of Computer Science and Engineering, Texas A&M University,
College Station, TX 77843, USA
e-mail: caverlee@cse.tamu.edu

K. Lee
Department of Computer Science, Utah State University, Logan, UT 84322, USA
e-mail: kyumin.lee@usu.edu

S.A. Matei et al. (eds.), *Transparency in Social Media*,
Computational Social Sciences, DOI 10.1007/978-3-319-18552-1_4

either online communities (56 %) or search engines (33 %). Unfortunately, little is known about Weaponized Crowdsourcing as it manifests in existing systems, nor what are the ramifications on the design and operation of emerging socio-technical systems. Hence, this chapter shall focus on key research questions related to Weaponized Crowdsourcing as well as outline the potential of building new preventative frameworks for maintaining the information quality and integrity of online communities in the face of this rising challenge.

2 Background

In a crowdsourcing marketplace like Amazon Mechanical Turk, a participant can be a requester: one who posts a task description and recruits workers to solve this task; a worker: one who performs a task and is typically compensated for this work; or both a requester and a worker. These tasks are usually difficult or computationally expensive for computers to solve, but relatively easy for humans. In many crowdsourcing marketplaces, complex tasks are typically broken down into simpler tasks that can be completed by an individual worker in a reasonable amount of time. For example, validating the quality of a transcribed script from an audio source (as in the case of using crowd workers to construct subtitles for a previously un-subtitled video) may be assigned to multiple, overlapping workers who tackle parts of the task: an individual worker may transcribe a 10-second clip; other workers may repeat this work or verify the quality of this work; eventually, the full-time transcription may then be completed and given to a final worker (or collection of workers) to validate. Workers in these crowdsourcing marketplaces are often cheaply paid and treated as interchangeable by requesters; and since workers are often drawn from the entire world, tasks may be completed at any time by a distributed workforce.

In light of these perceived benefits, we should note that a crowdsourcing marketplace is itself a social system that provides many of the advantages of social systems. That is, the reliance on users themselves to "maintain the community" can lead to many positive effects, including growth in the size and capabilities of the system, the emergence of recognized experts within the system (e.g., workers who are especially fast or precise, or have other desirable qualities), and the flexibility to tackle problems beyond the scope of the original system designers. And yet this relative openness and reliance on users to drive the system may lead to new risks and growing concerns. In particular, we highlight the challenge of *weaponized crowdsourcing*, in which malicious requesters misuse this openness to post tasks that spread malicious URLs in social media, form artificial grassroots campaigns, spread rumor and misinformation, and manipulate search engines. In the same vein, unethical workers will perform these tasks, often by propagating manipulated content to target sites such as social media sites, search engines, and review sites, resulting in the degradation of information quality and the integrity of these online communities.

To illustrate, Fig. 1 shows a typical workflow, wherein (1) a requester first posts one of these tasks (here, a “crowdturfing” task), (2) identifies the appropriate workers to complete this task, and (3) finally, these workers spread their misinformation in a target venue like a social network, a forum/review site, a search engine, or blog. Figure 2 shows an example of a crowdturfing task description that we sampled from the crowdsourcing platform Microworkers.com. This task requires workers to have at least 50 Twitter followers, search for a certain keyword on Google, and then click on a website in the search results. In addition, it requires the workers to retweet an article in the website to Twitter. This task targets not only a search engine but also a social media site, hoping to boost the target website’s rank by artificially manipulating both a search engine and a social network. At the time of our collection, 222 workers had completed this task for $0.60 per task completion.

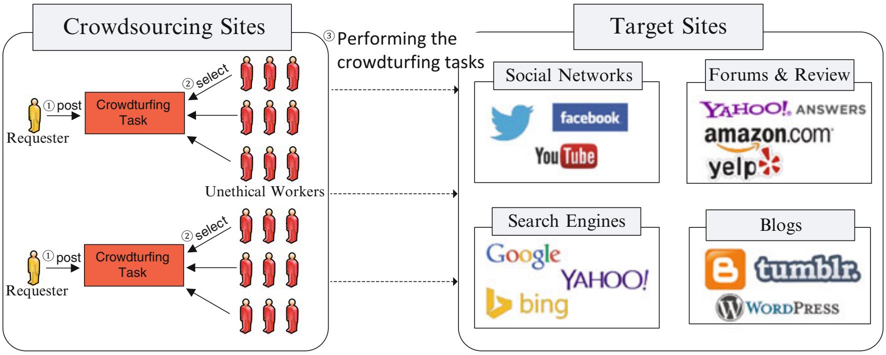
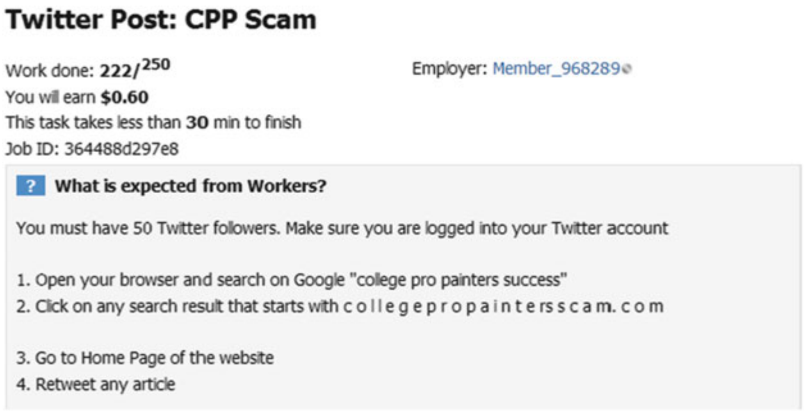


Fig. 1 The interactions between malicious requesters and unethical workers



Twitter Post: CPP Scam

Work done: **222/**250

Employer: Member_968289

You will earn **$0.60**

This task takes less than **30** min to finish

Job ID: 364488d297e8

What is expected from Workers?

You must have 50 Twitter followers. Make sure you are logged into your Twitter account

1. Open your browser and search on Google "college pro painters success"
2. Click on any search result that starts with collegepropaintersscam.com

3. Go to Home Page of the website
4. Retweet any article

Fig. 2 A crowdturfing task description posted to Microworkers.com

3 Weaponized Crowdsourcing: An Investigation

In this section, we investigate the emerging threat of weaponized crowdsourcing through a multi-part analysis. We sample and report on tasks from existing crowdsourcing marketplaces, characterize the market size, present an initial categorization of the types of campaigns, and investigate the demographics of both malicious requesters and unethical workers.

3.1 Datasets Collected from Crowdsourcing Sites

In order to conduct our analysis about weaponized crowdsourcing, we collected 505 campaigns totaling 63,042 tasks by crawling three popular Western crowdsourcing sites that host clear examples of crowdturfing campaigns: Microworkers.com, ShortTask.com, and Rapidworkers.com during a span of 2 months in 2012. Almost all of the campaigns in these sites are crowdturfing campaigns, and these sites are active in terms of the number of new campaigns being posted. Note that even though Amazon Mechanical Turk is one of the most popular crowdsourcing sites, we excluded it in our study because it has only a small number of crowdturfing campaigns and its terms of service officially prohibits the posting of crowdturfing campaigns. Perhaps surprisingly, Microworkers.com is ranked by Alexa.com at the 4,699th most popular website while Amazon Mechanical Turk is ranked 7,173. We additionally collected 89,667 campaign descriptions and 31,021 corresponding user profiles between July and August 2013 from Fiverr.com, a global microtask marketplace that as of April 2014 is the 130th most visited site in the world according to Alexa (2014).

3.2 Market Size of Weaponized Crowdsourcing

To analyze the market size of crowdturfing campaigns in Microworkers.com, We collected 144 requesters' profiles and 4,012 workers' profiles—where all campaigns in our sample data are crowdturfing tasks and other researchers have found that 89 % of campaigns hosted at Microworkers.com are indeed crowdturfing tasks (Wang et al., 2012).

The 4,012 workers have completed 2,962,897 tasks and earned $467,453 so far, which suggests the entirety of the crowdturfing market is substantial. Interestingly, the average price per task is higher on a crowdturfing site (for Microworkers.com, the average is $0.51) than on the legitimate Amazon Mechanical Turk where 90 percent of all tasks pay less than $0.10 (Ipeirotis, 2010).

Table 1 presents the maximum, average, median and minimum number of tasks done, how much they have earned, and the account longevity for the sampled

Table 1 Characteristics of Crowdturfing Workers in Microworkers.com

	# Of tasks	Total earned ($)	Longevity (day)
Max	24,016	3,699	1,215
Avg	738	117	368
Median	166	23	320
Min	10	1	5

Table 2 Characteristics of Crowdturfing Requesters in Microworkers.com

	# Of campaigns	# Of paid tasks	Longevity (day)
Max	4,137	455,994	1,091
Avg	68	7,030	329
Median	7	306	259
Min	1	0	3

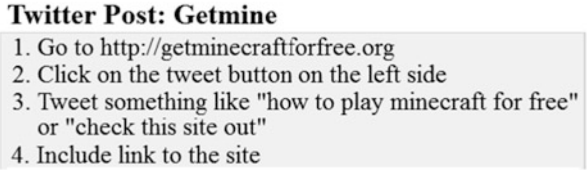
workers. We observe that there are professional workers who have earned reasonable money from the site to survive. For example, a user who earned $3,699 for slightly more than 3 years (1,215 days) lives in Bangladesh where the GNI (Gross National Income) per capita is $770 in 2011 as estimated by the World Bank TradingEconomics (2011). Surprisingly, she has earned even more money per year ($1,120) than the average income per year ($770) of a person in Bangladesh.

The requesters' profile information reveals their account longevity, number of paid tasks and expense/cost for campaigns. As shown in Table 2, many workers have created multiple campaigns with lots of tasks (on an average—68 campaigns and 7,030 paid tasks). The most active requester in our dataset initiated 4,137 campaigns associated with 455,994 paid tasks. In other words, he has spent a quarter million dollar ($232,557)—again a task costs $0.51 on an average. In total, 144 requesters have created 9720 campaigns with 1,012,333 tasks and have paid a half million dollars ($516,289). This sample analysis shows us how the dark market is big enough to tempt users from developing countries to become workers.

3.3 Types of Campaigns

We next analyze types of crowdturfing campaigns to understand the tactics of the requesters. Hence, we first manually grouped the 505 campaigns collected from Microworkers.com, ShortTask.com, and Rapidworkers.com into the following five categories:

- **Social Media Manipulation [56 %]:** The most popular campaigns target social media. Example campaigns request workers to spread a meme through social media sites such as Twitter, click the "like" button of a specific Facebook profile/ product page, bookmark a webpage on Stumbleupon, answer a question with a



Twitter Post: Getmine

1. Go to http://getminecraftforfree.org
2. Click on the tweet button on the left side
3. Tweet something like "how to play minecraft for free" or "check this site out"
4. Include link to the site

Fig. 3 An example social media manipulation campaign

link on Yahoo! Answers, write a review for a product at Amazon.com, or write an article on a personal blog. An example campaign is shown in Fig. 3, where workers are requested to post a tweet including a specific URL.

- **Sign Up [26 %]:** Requesters ask workers to sign up on a website for several reasons, for example to increase the user pool, to harvest user information like name and email, and to promote advertisements.
- **Search Engine Spamming [7 %]:** For this type of campaign, requesters seek to increase the visibility of a particular web page by creating artificial clicks, which are typically interpreted by major search engines as a signal of page quality. A typical task requires a worker to search for a specified keyword on a major search engine (like Google or Bing). The workers should then scan through the search engine results and click on the specified link (which is affiliated with the campaign's requester), towards increasing the number of clicks on the page and ultimately increasing the rank of the page in future searches, as shown in Fig. 2.
- **Vote Stuffing [4 %]:** Requesters ask workers to cast votes. In one example, the requester asked workers to vote for "Tommy Marsh and Bad Dog" to get the best blue band award in the Ventura County Music Awards (which the band ended up winning!).
- **Miscellany [7 %]:** Finally, a number of campaigns engaged in some other activity: for example, some requested workers to download, install, and rate a particular software package; others requested workers to participate in a survey or join an online game.

Next, we also analyzed 121 crowdturfing campaigns randomly sampled from Fiverr.com, and manually grouped them into three categories:

- **Social Media Targeting Campaigns [54 %]:** These crowdturfing campaigns targeted social media sites such as Facebook, Twitter and Youtube. The main purpose of these campaigns are to artificially increase number of friends or followers on these sites, promote pre-selected messages or URLs, and increase the number of views associated with requesters' videos. The requesters expect these manipulations to result in more effective information propagation, higher conversion rates, and positive social signals for their web pages and products.
- **Search Engine Targeting Campaigns [38 %]:** These campaigns targeted search engines by artificially creating backlinks for a targeted site. This is a traditional attack against search engines. However, instead of creating backlinks on their own, the requesters take advantage of workers to create a large number of

backlinks so that the targeted page will receive a higher PageRank score (and have a better chance of ranking at the top of search results). Interestingly, a worker (also called a seller in Fiverr.com) has earned $3 million for helping running search engine targeting campaigns with 100 % positive ratings and more than 47,000 positive comments from requesters who hired the worker. This fact indicates that the search engine targeting campaigns are popular and profitable.

- **User Traffic Targeting Campaigns [8 %]:** The last campaigns aimed to get user traffic to a targeted site. Workers generated user traffic (visitors) for a pre-selected website or web page. With higher traffic, the requesters hope to abuse Google AdSense, which provides advertisements on each requester's web page, when the visitors click the advertisements. Another goal of these campaigns is for the visitors to purchase products from the pre-selected page.

From the analysis of types of crowdturfing campaigns, we can see that most existing crowdturfing campaigns have targeted social media sites and search engines, which raises natural concerns about the information quality and community trust of these systems.

3.4 Countries of Requesters and Workers

Next we analyze where requesters and workers were from in Microworkers.com and Fiverr.com. Do workers and requesters have different country distributions? Can we observe different country distributions of requesters and workers who were involved in crowdturfing campaigns in Microworkers.com and Fiverr.com?

To answer these research questions, we first analyze the countries of workers and requesters in Microworkers.com. From the 4,012 workers' profile information in Microworkers.com, we found that they are from 75 countries. Especially, 83 % of the workers are from the top-10 countries as shown in Fig. 4a. An interesting

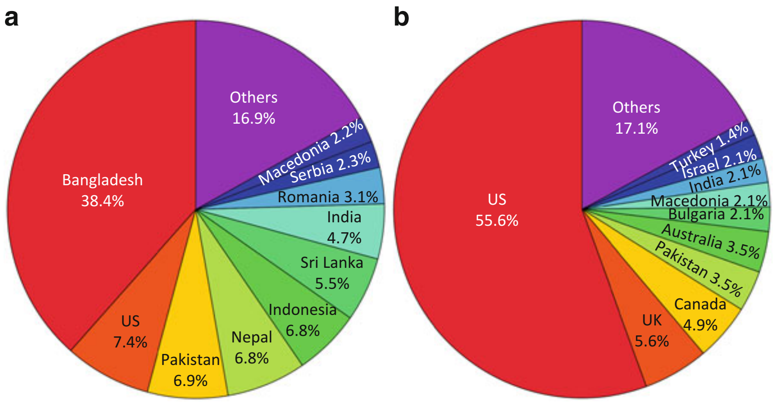


Fig. 4 Top 10 countries of workers and requesters of crowdturfing campaigns in Microworkers.com. (**a**) Workers, (**b**) requesters

observation is that a major portion of the workers in Microworkers.com are from Bangladesh—where 38 % workers (1,539 workers) come from—whereas in Amazon Mechanical Turk over 90 % workers are from the United States and India Ross, Irani, Silberman, Zaldivar, and Tomlinson (2010).

However, requesters in Microworkers.com have a different country distribution. We found that the requesters are from 31 countries. Interestingly, 55 % of the requesters are from the United States, and 70 % of the requesters are from the English-speaking countries: United States, UK, Canada, and Australia. Figure 4b shows the top-10 countries which have the highest portion of requesters. We can see an imbalance between the country of origin of requesters and of the workers, but that the ultimate goal is to propagate artificial content through the English-speaking web.

Next, we analyze countries of workers and requesters in Fiverr.com, and compare their country distribution with country distribution of workers and requesters in Microworkers.com. Interestingly, the most frequent workers who performed crowdturfing tasks were from the United States (35.8 %) as shown in Fig. 5a. The next largest group of workers is from India (10.5 %), followed by Bangladesh (6.5 %) and the United Kingdom (5.9 %). Overall, the majority of workers (52 %) were from western countries. This distribution is very different from country distribution of workers in Microworkers.com in which the most frequent workers were from Bangladesh. This observation might imply that Fiverr.com is more attractive than Microworkers.com for U.S. residents since a worker in Fiverr.com earns higher income (at least $5 per task) than a worker in Microworkers.com (average $0.50 per task). The country distribution of requesters in Fiverr.com (as shown in Fig. 5b) is similar with a country distribution of requesters in Microworkers.com, in which the majority of requesters were from English-speaking countries.

So far, we have investigated the weaponized crowdsourcing market size, examined the distribution of tasks on two platforms, and seen how these platforms attract both workers and requesters from around the world to target successful social and web communities.

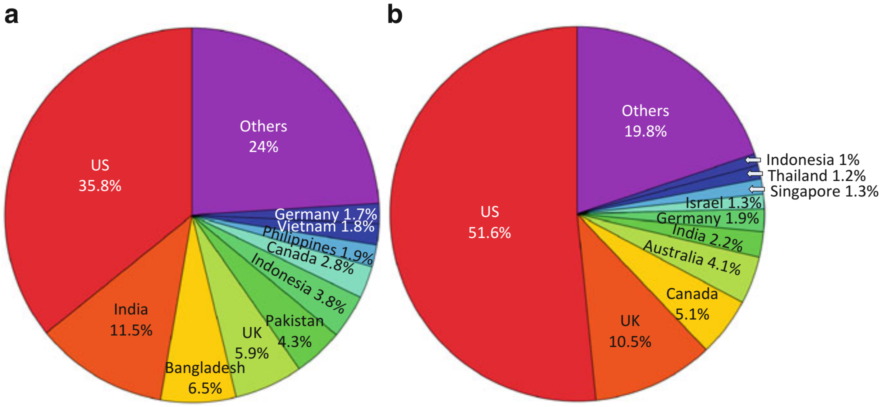


Fig. 5 Top 10 countries of workers and requesters of crowdturfing campaigns in Fiverr.com. (**a**) Workers, (**b**) requesters

4 Preventive Approaches

Given the scale and reach of existing weaponized crowdsourcing marketplaces and concerns over how they may grow in the future, we turn in this section to a discussion of possible preventative approaches for mitigating their impact on socio-technical systems. Our goal is to highlight approaches to detect and prevent the weaponized crowdsourcing problem. Specifically, we highlight three approaches: (i) an approach to detect crowdturfing tasks at the source (in the crowdsourcing platform itself); (ii) an approach to detect accounts of crowd workers who performed crowdturfing tasks in a target site (by looking at the impacts of these tasks in their target); and (iii) a crowdsourcing approach that aims to use the crowd itself to monitor and police itself. Then we turn to a discussion of future steps toward improving our defenses against weaponized crowdsourcing.

4.1 Automatic Crowdturfing Task Detection

One way to solve the crowdturfing problem is to automatically detect and delete crowdturfing tasks to prevent workers from performing the crowdturfing tasks. To measure whether automatically detecting crowdturfing tasks is possible, we present here a prototype crowdturfing task detection classifier.

First, we randomly sampled 1,550 distinct tasks from Fiverr.com and manually labeled them as either a legitimate task or a crowdturfing task. As we described in Sect. 3.3, we found that 121 out of 1,550 tasks were crowdturfing tasks. This labeled dataset was converted to feature values to train and test our SVM-based classifier. Our feature set consists of the title of a task, the task's description, a top level category, a second level category (each task at Fiverr.com is categorized to a top level and then a second level—e.g., "online marketing" as the top level and "social marketing" as the second level), ratings associated with a task, the number of votes for a task, a task's longevity and so on (detailed information can be found in Lee, Webb, & Ge, 2014). For the title and job description of a task, we converted these texts into bag-of-word models in which each distinct word becomes a feature. We also used *tf-idf* to measure values for these text features.

Then, we trained and tested our SVM-based classifier with tenfold cross-validation. Its classification result is shown in Table 3, where we can see a 97.35 % accuracy, 0.974 F_1, 0.008 false positive rate (FPR), and 0.248 false negative rate (FNR). This positive result shows that our classification approach works well.

Table 3 SVM-based classification result

Accuracy	F_1	FPR	FNR
97.35 %	0.974	0.008	0.248

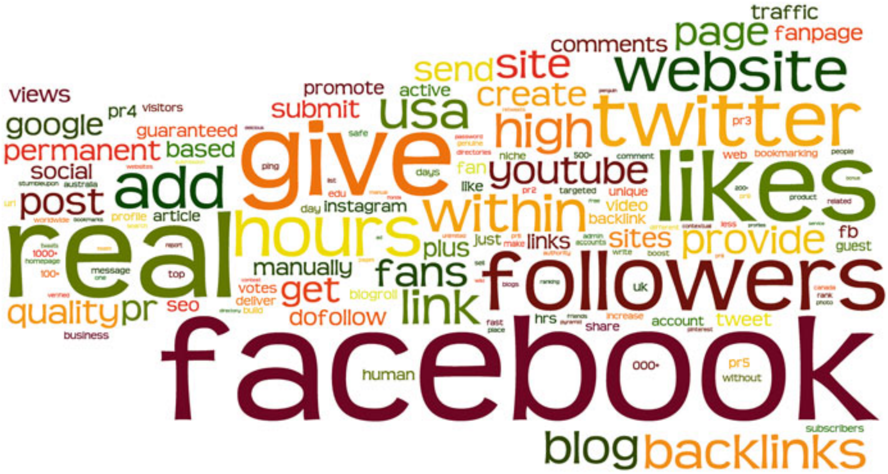


Fig. 6 Word cloud of crowdturfing tasks

We also applied this classifier to a larger testing set containing 87,818 tasks. In this experiment, the 1,550 tasks were used as a training. We built the SVM-based classifier with the training set and predicted class labels of the tasks in the testing set. 19,904 of the 87,818 tasks were predicted as crowdturfing tasks. To verify whether the predicted 19,904 crowdturfing tasks are real crowdturfing tasks, we manually scanned the titles of all of these tasks and confirmed that our approach worked well.

To understand and visualize what terms crowdturfing tasks often contain, we generated a word cloud of titles for these 19,904 crowdturfing tasks. First, we extracted the titles of the tasks and tokenized them to generate unigrams. Then, we removed stop words. Figure 6 shows the word cloud of crowdturfing tasks. The most popular terms are online social network names (e.g., Facebook, Twitter, and YouTube), targeted goals for the online social networks (e.g., likes and followers), and search engine related terms (e.g., backlinks, website, and Google). This word cloud also helps confirm that our classifier accurately identified crowdturfing tasks.

The experimental results confirm that automatically detecting crowdturfing tasks are possible. We expect this approach would filter crowdturfing tasks before workers take the jobs.

4.2 Tracking Manipulated Content and Detecting Workers in Social Media

Another way to solve the crowdturfing problem is to detect crowd workers' accounts in target sites. By linking manipulated content such as URLs and message templates to a target site, we would identify crowd workers' accounts in the target site. By learning these crowd workers' behaviors in the target site, we may automatically detect accounts of crowd workers who have performed crowdturfing tasks. To test this possibility, we selected 65 campaigns, which targeted Twitter, from 505 campaigns collected from Microworkers.com, ShortTask.com, and Rapidworkers.com. There were two types of Twitter related crowdturfing campaigns—campaigns which ask to post a tweet and the ones which ask to follow a user.

- **Tweeting about a link:** These tasks ask the Twitter workers to post a tweet including a specific URL (as in the example in Fig. 3). The objective is to spread a URL to other Twitter users, and thereby increase the number of clicks on the URL.
- **Following a Twitter user:** The second task type requires a Twitter worker to follow a requester's Twitter account. These campaigns can increase the visibility of the requester's account (for targeting larger future audiences) as well as impacting link analysis algorithms (like PageRank and HITS) used in Twitter search or in general Web search engines that incorporate linkage relationships in social media.

Next we tracked the Twitter accounts who participated in these campaigns. For campaigns of the first type, we used the Twitter search API to find all Twitter users who had posted the URL. For campaigns of the second type, we identified all users who had followed the requester's Twitter account. In total, we identified 2,864 Twitter workers. For these workers, we additionally collected their Twitter profile information, most recent 200 tweets, and social relationships (followings and followers).

In order to compare how these workers' properties are different from non-workers, we randomly sampled 10,000 Twitter users. Since we have no guarantees that these sampled users are indeed non-workers, we monitored the accounts for 1 month to see if they were still active and not suspended by Twitter. After 1 month, we found that 9,878 users were still active. In addition, we randomly selected 200 users out of the 9,878 users and manually checked their profiles, and found that only 6 out of 200 users seemed suspicious. Based on these verifications, we labeled the 9,878 users as non-workers. Even though there is a chance of a false positive in the non-worker set, the results of any analysis should give us, at worst, a lower bound since the introduction of possible noise would only degrade our results.

To build a crowd worker detection classifier, we converted the dataset containing information of 2,864 workers and 9,878 non-workers to feature values. Our features consist of four feature groups:

- **User Demographics (UD):** features extracted from descriptive information about a user and his account.
- **User Friendship Networks (UFN):** features extracted from friendship information such as the number of followings and followers.
- **User Activity (UA):** features representing posting activities.
- **User Content (UC):** features extracted from posted tweets.

Table 4 Features

Group	Feature
UD	The length of the screen name
UD	The length of description
UD	The longevity of the account
UD	Has description in profile
UD	Has URL in profile
UFN	The number of followings
UFN	The number of followers
UFN	The ratio of the number of followings and followers
UFN	The percentage of bidirectional friends: $\frac{\lvert followings \cap followers \rvert}{\lvert followings \rvert}$ and $\frac{\lvert followings \cap followers \rvert}{\lvert followers \rvert}$
UA	The number of posted tweets
UA	The number of posted tweets per day
UA	| links | in tweets / | tweets |
UA	| hashtags | in tweets / | tweets |
UA	| @username | in tweets / | tweets |
UA	| rt | in tweets / | tweets |
UA	| tweets | / | recent days |
UA	| links | in tweets / | recent days |
UA	| hashtags | in tweets / | recent days |
UA	| @username | in tweets / | recent days |
UA	| rt | in tweets in tweets / | recent days |
UA	| links | in RT tweets / | RT tweets |
UC	The average content similarity over all pairs of tweets posted: $\frac{\sum similarity(a,b)}{\lvert \text{set of pairs in tweets} \rvert}$, where $a, b \in$ set of pairs in tweets
UC	The ZIP compression ratio of posted tweets: $\frac{uncompressed\ size\ of\ tweets}{compressed\ size\ of\ tweets}$
UC	68 LIWC features (Pennebaker, Francis, & Booth, 2001) which are Total Pronouns, 1st Person Singular, 1st Person Plural, 1st Person, 2nd Person, 3rd Person, Negation, Assent, Articles, Prepositions, Numbers, Affect, Positive Emotions, Positive Feelings, Optimism, Negative Emotions, Anxiety, Anger, Sadness, Cognitive Processes, Causation, Insight, Discrepancy, Inhibition, Tentative, Certainty, Sensory Processes, Seeing, Hearing, Touch, Social Processes, Communication, Other References to People, Friends, Family, Humans, Time, Past Tense Verb, Present Tense Verb, Future, Space, Up, Down, Inclusive, Exclusive, Motion, Occupation, School, Job/Work, Achievement, Leisure, Home, Sports, TV/Movies, Music, Money, Metaphysical States, Religion, Death, Physical States, Body States, Sexual, Eating, Sleeping, Grooming, Swearing, Nonfluencies, and Fillers

From the four groups, we generated a total 92 features as shown in Table 4 (detailed information can be found in Lee, Tamilarasan, et al., 2013).

Using tenfold cross-validation approach and these feature groups, we tested 30 classification algorithms using the Weka machine learning toolkit (Witten & Frank, 2005). To test which classification algorithm returns the highest accuracy, we ran over 30 classification algorithms such as Naive Bayes, Logistic Regression and SMO (SVM) with the default setting. Their accuracies ranges from 86 to 91 %.

Table 5 Worker detection: results

Classifier	Accuracy	F_1	AUC	FPR	FNR
Random Forest	93.26 %	0.966	0.955	0.036	0.174

Tree-based classifiers showed the highest accuracy results. In particular, Random Forest produced the highest accuracy which was 91.85 %. By changing input parameter values of Random Forest, we achieved 93.26 % accuracy and 0.932 F_1 as shown in Table 5.

The experimental results confirm that we can automatically detect accounts of crowd workers who performed crowdturfing tasks.

4.3 Crowdsourced Mitigation

Another possible approach is to mobilize the crowd itself to mitigate the threat of weaponized crowdsourcing. But how can a crowd be organized to police itself? In one direction, we could hire crowd workers whose job is to verify whether a task is crowdturfing or not. This approach can be combined with the above approaches. For example, a crowdturfing task detector could give us a probabilistic assessment of each task (e.g., task A would be a crowdturfing task with 80 % probability). Since sometimes the detector may give us some false negatives, predicted crowdturfing tasks with a low probability would be passed to crowd workers and verified to build a more accurate crowdturfing detection system. A similar work to detect social spammers by crowd workers was studied by Wang et al. (2013).

4.4 Discussion

So far, we have introduced several *algorithmic* approaches for maintaining the information quality and integrity of online communities in the face of weaponized crowdsourcing. We now turn to a forward-looking discussion of other socio-technical approaches including collaboration among crowdsourcing service providers, target companies (e.g., social media and search engine companies), and the government.

First, we suggest creating and maintaining a common repository where employees of crowdsourcing sites and researchers store crowdturfing task descriptions containing manipulated content (e.g., URLs, template messages). Email and web service providers already maintain blacklists to store malicious web page URLs for spam, phishing and malware software distribution. A new repository for crowdturfing tasks would be helpful for employees of crowdsourcing and target sites, and for researchers to actively detect and prevent crowdturfing tasks, manipulated content, and participants.

Second, we have to think of how to increase the cost of running crowdturfing campaigns and how to discourage workers from participating in these campaigns. For example, we could imagine forfeiting malicious requesters' credits and blocking their IP addresses, as well as suspending unethical workers' accounts and blocking their IP addresses. When a user creates an account in a crowdsourcing site, we could require providing an email account and passing a Captcha so that we can delay these malicious requesters and workers from creating accounts and discourage running and participating in crowdturfing campaigns.

An interesting observation that we learned from this work is there are several crowdsourcing sites where almost all tasks are crowdturfing tasks. These crowdsourcing site providers intentionally do not prohibit posting crowdturfing tasks because these providers earn commission (about 20 %) from requesters. In addition, as we mentioned in Sect. 3.2, a crowdturfing task is five times more expensive than a legitimate task, further encouraging these crowdsourcing platforms to allow these crowdturfing tasks. To solve this problem, another potential effort is for governments or specialized organizations to start monitoring crowdsourcing sites for these weaponized crowdturfing tasks, and then to advocate for a strong response (e.g., bringing public pressure to shut down these sites).

5 Conclusion

In this chapter, we have highlighted the challenges presented by weaponized crowdsourcing and begun a discussion of potential countermeasures. As crowdsourcing platforms and systems continue to grow in complexity, variety, and reach, we can naturally anticipate the continued challenge and maturation of threats posed by weaponized crowdsourcing. Moving forward, we believe that weaponized crowdsourcing research is poised to make major breakthroughs in the years to come due to the growing interest and collaboration of researchers and practitioners across disciplines toward improving the transparency and trust of social media and online interactions.

Contributions Portions of this chapter are based on work that appeared in the 2013 and 2014 International AAAI Conference on Weblogs and Social Media (ICWSM) (Lee, Tamilarasan, et al., 2013; Lee et al., 2014).

References

Alexa (2014). Fiverr.com site info—alexa. http://www.alexa.com/siteinfo/fiverr.com.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. In *WWW*.

Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., & Zhao, B. Y. (2010). Detecting and characterizing social spam campaigns. In *Proceedings of the 10th annual conference on Internet measurement (IMC)*.

Grier, C., Thomas, K., Paxson, V., & Zhang, M. (2010). @spam: The underground on 140 characters or less. In *CCS*.

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013). Faking sandy: Characterizing and identifying fake images on twitter during hurricane sandy. In *WWW Companion*.

Ipeirotis, P. G. (2010). Analyzing the amazon mechanical turk marketplace. In *XRDS*, (Vol. 17, pp. 16–21).

Lee, K., Caverlee, J., Cheng, Z., & Sui, D. Z. (2013). Campaign extraction from social media. *ACM Transactions on Intelligent Systems and Technology, 5*, 9:1–9:28.

Lee, K., Tamilarasan, P., & Caverlee, J. (2013). Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media. In *ICWSM*.

Lee, K., Webb, S., & Ge, H. (2014).The dark side of micro-task marketplaces: Characterizing fiverr and automatically detecting crowdturfing. In *ICWSM*.

Lee, S., & Kim, J. (2012). Warningbird: Detecting suspicious urls in twitter stream. In *NDSS*.

Pennebaker, J., Francis, M., & Booth, R. (2001). *Linguistic inquiry and word count*. Mahwah: Erlbaum Publishers.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI Extended Abstracts on Human Factors in Computing Systems*.

TradingEconomics (2011). Gni per capita; atlas method (us dollar) in bangladesh. http://www.tradingeconomics.com/bangladesh/gni-per-capita-atlas-method-us-dollar-wb-data.html.

Wang, G., Mohanlal, M., Wilson, C., Wang, X., Metzger, M. J., Zheng, H., et al. (2013). Social turing tests: Crowdsourcing sybil detection. In *NDSS*.

Wang, G., Wilson, C., Zhao, X., Zhu, Y., Mohanlal, M., Zheng, H., et al. (2012). Serf and turf: crowdturfing for fun and profit. In *WWW*.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, 2nd ed. New York: Morgan Kaufmann.