Exploitation in Human Computation Systems

James Caverlee

Introduction

Much of this handbook has been devoted to the positive potential and possibilities unlocked by human computation systems. From specialized systems like Ushahidi (for crisis mapping), Foldit (for protein folding) and Duolingo (for foreign language learning and translation) to general-purpose crowdsourcing platforms like Amazon Mechanical Turk and Crowdflower—these systems have shown the effectiveness of intelligently organizing large numbers of people, and suggest a rich future for next generation human computation systems.

In this chapter, we turn our sights to the negative aspects of these systems. How may participants be exploited by human computation systems? How can these systems be used as a means to exploit other populations? What are the existing types of exploits and what types of exploits does the future hold? Beyond characterizing the threat horizon, we also consider efforts toward detecting exploits in human computation systems. And what are steps that can be taken toward mitigating the risk as these systems continue to mature?

Exploitation Within a Human Computation System

In this section, we present opportunities for exploitation within a human computation system. We consider exploits that target workers (who actually perform jobs), exploits that target requesters (who solicit jobs), and finally exploits that target the system as a whole. This taxonomy is intended as an initial organization of some of the exploits facing human computation systems, and should not be considered comprehensive.

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

P. Michelucci (ed.), *Handbook of Human Computation*,
DOI 10.1007/978-1-4614-8806-4_68,

Exploits Targeting Workers

Human computation systems rely on the support of *workers* who are tasked with supporting the overall efforts of the system.

Misrepresentation of the task. In many human computation systems, the overall effort is subdivided into smaller chunks that may be handled by individual workers. For example, a system to automatically recognize animals captured on video may provide each individual worker with access to only a few key frames from a handful of videos. As a result, the worker may have only incomplete knowledge of the ultimate goal of the task. In many settings, this incomplete knowledge is uncontroversial.

However, this compartmentalization of task knowledge may lead to workers agreeing to participate in human computation systems where the overall effort is contrary to the worker's moral, ethical, or religious grounding. For example, Jonathan Zittrain characterized this exploit as such:

> You might synthesize a new chemical that winds up being used as a poison or in a bomb. Iran's leaders could ask Turkers to cross-reference the faces of the nation's 72 million citizens with those of photographed demonstrators. Based on Mechanical Turk's current rates, Repression 2.0 would cost a mere $17,000 per protester. (Zittrain 2009)

Another example of abusing workers morals via task misrepresentation would be saying you are tracking elephant movement supposedly for conservation, but the data is used by poachers. In this way, workers may become cogs in a machine that works counter to their own interests.

Exposure to unwanted risks. Even for tasks that are agreeable to a worker, a worker in a human computation system may be exposed to risks that go beyond their reasonable expectations. In one direction, a worker may be exposed to disturbing content (say, via an image labeling task). In a separate direction, a worker may encounter misinformation spread through an otherwise legitimate task. For example, a worker may be asked to label blog posts as containing evidence of propaganda or not; through the labeling process, the worker may encounter deliberately placed misinformation designed to change the worker's perceptions of a particular candidate or political issue (e.g., climate change). Such a risk is similar to "push polling" in traditional political campaign surveys whereby a polling question is deliberately constructed to persuade (or even mislead) a respondent.

In addition to the cognitive risks of exposure, workers and their computing systems may also be subject to spam, malware, and phishing (Jagatic et al. 2007) attacks that have shown a remarkable ability to migrate to emerging systems. From email to Web to social media, and eventually to human computation systems, malicious users have shown great ability to target new populations.

Privacy leakage. Workers in a human computation system may also subject themselves to potential loss of privacy. A recent study has found that Amazon Mechanical Turk—designed to be an anonymous system—leaks private information of workers by using a single unique identifier for all Amazon accounts (Lease et al. 2013). In

this way, a worker's anonymous Mechanical Turk account can be linked to the same worker's Amazon profile page, which could reveal personally identifying information. Beyond the direct negative consequences of privacy leakage (e.g., loss of user anonymity, targeted attacks on individuals de-identified), a worker's willingness to participate in human computation systems may be limited if there are perceived risks of privacy leakage.

Unsatisfactory compensation. The final exploit has been widely recognized as a potential threat in the increasingly globalized virtual workforce enabled by human computation systems (Ross et al. 2010). By drawing on workers from low income countries, there is the potential for exploitation of disadvantaged workers.

Exploits Targeting Requesters

On the other hand, there are threats to the requesters in human computation systems (or to the overall operators of the system).

Competitive disruption. In the 2009 DARPA Red Balloon Challenge, the winning MIT team reported that some participants deliberately falsified balloon sightings, whether to disrupt the overall functioning of the overall requester goal (find all of the balloons) or to disrupt the balloon sightings of competitor teams (Tang et al. 2011). In this way, groups of workers within a system or a competitor system itself may negatively impact the functioning of a target system by delaying task completion time, by degrading the quality of work being done (say, through deliberately inserting misinformation), and by adding uncertainty to the overall reliability of the system.

Poor quality work. One of the key concerns for requesters using existing systems like Amazon Mechanical Turk's crowdsourcing marketplace is the quality of work provided by workers. It is possible that the quality of work provided by workers is of lower quality than advertised by the human computation system: e.g., workers, regardless of incentivization scheme, may choose to complete as many tasks as possible while exerting little effort. For example, in a task that is answered using multi-choice options, the worker might randomly select answers, or in case of tasks that require answering verbosely (review of a product, comparison between two products) the worker might use generic answers or answers off of Internet to complete the task quickly. In addition to this, another reason for poor quality of work on a human computation system could be because of the "one size fits all" expectation that requesters have of the system. The requester might observe a mismatch in worker skills between what the system can provide and what they are expecting. For example, a human computation system might mostly have English speaking workers, but a requesters task might need knowledge of Chinese that the system might not be able to satisfy. Existing systems (like Amazon Mechanical Turk) do include capabilities to track worker performance across tasks, to filter participants by native language, and other "checks and balances" to overcome some of these quality

issues. However, as human computation systems increase in variety and capabilities, maintaining quality work will be a fundamental challenge.

Privacy leakage. As in the case of workers on a human computational system, a requester's privacy may be leaked. This can be either due to the design of the computation system itself, for example, a work requester's Mechanical Turk account being linked to his Amazon profile page, or it could be because of the information that the requester inadvertently added to the task like his email, company he works for, and so on. A requester's privacy leakage could result in reduced quality in work. For example on Amazon's Mechanical Turk, the requester has the right to pay or not pay a worker for the task completed based on the worker's quality of work. A worker who knows requester's details could answer the task in a biased way (praise requester's company, prefer requester's approach while two items are compared) so as to impress the requester without actually doing the task correctly. Also in case the requester rejects worker's work then a worker who knows requester's contact details can get in touch with him requesting the details or even threaten him.

Exploits Targeting the System Itself

Finally, human computation systems themselves may come under threat by external parties interested in degrading the quality of online information and threatening the usefulness of these systems. Traditional denial of service, spam, and other targeted attacks can be modified to disrupt the reliability, quality, and timeliness of human computation systems.

Exploits Targeting External Populations

In this section, we consider opportunities for malicious users to leverage human computation systems to target external (outside of the system) populations. We couple this treatment with a study of the prevalence of one type of exploit (crowdturfing), and consider additional exploits.

Crowdturfing

One growing threat is the emergence of "crowdturfing" (crowdsourcing + astroturfing), whereby masses of cheaply paid shills can be organized to spread malicious URLs in social media, form artificial grassroots campaigns ("astroturf"), and

manipulate search engines. One example is the development of sites like SubvertAndProfit (www.subvertandprofit.com), which claims to have access to "25,000 users who earn money by viewing, voting, fanning, rating, or posting assigned tasks" across social media sites. These campaigns are being launched from commercial crowdsourcing sites, potentially leading to the commoditization of large-scale turfing campaigns. In a recent study of the two largest Chinese crowdsourcing sites Zhubajie and Sandaha, Wang et al. (2012) found that ~90 % of all tasks were for crowdturfing.

Evidence of Crowdturfing

To illustrate the impact of crowdturfing, we report here a brief study of 505 campaigns collected from 3 popular Western crowdsourcing sites that host clear examples of crowdturfing campaigns: Microworkers.com, ShortTask.com, and Rapid-workers.com during a span of 2 months in 2012. Almost all campaigns in these sites are crowdturfing campaigns, and these sites are active in terms of number of new campaigns. 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. For the 505 sampled campaigns, each has multiple tasks, totaling 63,042 tasks. Based on a manual assignment, we found five major crowdturfing campaign types:

Social Media Manipulation (56 %). The most popular type of campaign targets 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 link on Yahoo! Answers, write a review for a product at Amazon.com, or write an article on a personal blog.

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, workers are asked to search for a certain keyword on a search engine, and then click the specified link (which is affiliated with the campaign's requester), toward increasing the rank of the page.

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.

Other Example Exploits

Propaganda. Crowdturfing can be leveraged for spreading misinformation and propaganda. For example, it has been recently reported that Vietnamese propaganda officials deployed 1,000 propagandists to engage in online discussions and post comments supporting the Communist Party's policies (BBC 2013). Similarly, the Chinese "Internet Water Army" can be hired to post positive comments for the government or commercial products, as well as disparage rivals (Wired 2010). Mass organized crowdturfers are also targeting popular services like iTunes (Gizmodo 2012) and attracting the attention of US intelligence operations (Guardian 2011).

Coordinated attacks. By exploiting collaboration to solve problems, newly engineered human computation systems could create novel ways of perpetrating crimes, acts of war, and other attacks. As illustration of this potential, in February 2013 a criminal syndicate infiltrated a credit card processing company, raised the withdrawal limits of ATM cards, and then distributed these ATM cards to dozens of participants around the world to simultaneously withdraw $45 million. Now imagine a similar attack coordinated via a human computation system whereby thousands of participants collaborate in a similar fashion. Beyond criminal activity, coordinated crowdsourced attacks could be used to decrypt passwords or launch cyber attacks on the computer systems of a country's adversaries. Perhaps even more troubling, a coordinated attack by a large group could mask their malicious behavior by acting collectively so that their influence on the system cannot be traced to a single aberrant individual.

Crowdsourced click manipulation. We have observed crowd workers leveraging human-powered crowdsourcing platforms to intentionally manipulate click patterns of URLs spread through social media to create conditions of artificial collective attention, in effect to create the illusion of collective attention toward increasing the population exposed to a malicious URL (say, by pushing the message containing such a URL into the day's trending topics on a system like Twitter) (Lee et al. 2013a).

Location-based deception. The rise of global-scale location sharing services (like Foursquare, and services supporting fine-grained location sharing like Instagram) allow users to connect in the physical world by revealing their footprints (typically via a "check-in" containing the user's current location that is shared through a social media service), leading to a host of positive opportunities. But these services can be misused to manipulate collective attention. In discussions with the Austin (Texas) Police Department, we have identified the threat of intentional deception through

the creation of fake "check-ins" around protests so that police response may be redirected to the wrong location.

Notice that these threats may have far reaching consequences, if successfully carried out. For example, during the recent Hurricane Sandy, several episodes of misinformation have led to confusion, errors, and slowed down humanitarian actions in affected zones, causing FEMA to formally address the issue (FEMA 2012; Meier 2012). For example, social media users posted fake storm images and spread misinformation that FEMA had run out of bottled water. Given the magnitude of the storms, FEMA has acknowledged the great role of social media as an effective means to quickly gain collective attention, but identified misinformation as a real threat to human lives.

Methods to Detect and Mitigate Exploits

Detecting exploits in human computation system is quite important, and the corresponding detection technique varies based on the type of exploit.

Reputation Systems. Many e-marketplaces and online communities use reputation systems to assess the quality of their members, including eBay, Amazon, and Digg, and reputation-based trust systems have received considerable research attention, e.g., Marti and Garcia-Molina (2006) and Resnick et al. (2000). These approaches aggregate community knowledge for evaluating the trustworthiness of participants. The benefits of reputation-based trust from a user's perspective include the ability to rate neighbors, a mechanism to reach out to the rest of the community. Along these lines, the recently proposed Turkopticon (Irani and Silberman 2013) is one such reputation system designed for human computation systems, in which workers on Amazon Mechanical Turk can rate interactions with requesters.

Policy-Based Approaches. Separately, exploits in human computation systems may be dealt with by rule-based or policy-oriented approaches. For obvious exploit like "workers getting poorly paid", it is intuitive just to compare the estimated average payment per hour for a task to the legal minimum wage rate, to decide whether workers are being exploited for lack appreciation of their efforts. In a more systematic manner, there has been some recent work on monitoring the quality of workers and their outputs. For example, Venetis and Garcia-Molina (2012) described two quality control mechanisms. The first mechanism repeats each task multiple times and combines the results from multiple users. The second mechanism defines a score for each worker and eliminates the work from users with low scores. Xia et al. (2012) provided a real-time quality control strategy for workers who evaluate the relevance of search engine results based on the combination of a qualification test of the workers (i.e., a question for which the requester already knows the answer) and the time spent on the actual task. The results are promising and these strategies facilitate reducing the number of bad workers.

Machine Learning. For more complicated exploits that target external populations, machine learning techniques can be applied. In one direction, the artifacts of a crowd powered targeting of an external population can be analyzed to develop machine learning models of the activities of the users who engaged in this activity. For example, by analyzing the social media artifacts of astroturf campaigns, researchers have developed methods to automatically detect crowd powered campaigns and the users engaged in these campaigns (Gao et al. 2010).

In a separate direction, since many current crowd turfing approaches target social media, researchers have proposed a framework for linking tasks (and their workers) on crowdsourcing sites to social media, by monitoring the activities of social media participants (Lee et al. 2013b). In this way, we can track the activities of crowdturfers in social media where their behavior, social network topology, and other cues may leak information about the underlying crowdturfing ecosystem. Based on this framework, researchers have identified the hidden information propagation structure connecting these workers in Twitter, which can reveal the implicit power structure of crowdturfers identified on crowdsourcing. Specifically, three classes of crowdturfers have been identified—professional workers, casual workers, and middlemen; based on statistical user models these users can be automatically differentiated from regular social media users.

Crowd-Based Mitigation. Finally, the crowd itself may be mobilized to mitigate exploits. How can a crowd be organized to police itself? How can a crowd detect the exploits within its own system and mitigate the impacts of exploits powered by other systems? In one direction, a crowd-powered monitoring system (akin to the Turkopticon) could be extended so that sub communities within the system validate the tasks within the system, towards reducing the opportunity of exploits to gain sufficient traction. Similarly, crowds could be deployed to monitor external communities (as on social media) for evidence of exploits; such a crowd-powered system could alert external communities of exploits and even roll-back negative actions (e.g., undoing Wikipedia vandalism). Of course such a crowd-checking-crowd system raises questions of "who watches the watchmen?" which we leave as an open and enduring question.

Summary

This chapter has presented a characterization of exploits that may target participants within human computation systems, as well as exploits that may target other populations. As crowd-powered systems continue to become more complex and of greater variety, we would expect a commensurate maturation of the exploit vectors, and (hopefully) of the technical and policy-oriented countermeasures to mitigating their impact.

References

BBC (2013) Vietnam admits deploying bloggers to support government. http://www.bbc.co.uk/news/world-asia-20982985, Jan 2013

FEMA (2012) Hurricane sandy: rumor control. http://www.fema.gov/hurricane-sandy-rumor-control

Gao H, Hu J, Wilson C, Li Z, Chen Y, Zhao BY (2010) Detecting and characterizing social spam campaigns. In: IMC, Melbourne

Gizmodo (2012) How a fake erotic fiction ebook hit the top 5 of itunes. http://gizmodo.com/5933169/how-a-fake-crowdsourced-erotic-ebook-hit-the-top-5-of-itunes, Aug 2012

Guardian T (2011) Revealed: US spy operation that manipulates social media. http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks, Mar 2011

Irani L, Silberman MS (2013) Turkopticon: interrupting worker invisibility in amazon mechanical turk. In: ACM SIGCHI conference on human factors in computing systems, Paris

Jagatic TN, Johnson NA, Jakobsson M, Menczer F (2007) Social phishing. Commun ACM 50(10):94–100

Lease et al. M (2013) Mechanical turk is not anonymous

Lee K, Kamath K, Caverlee J (2013a) Combating threats to collective attention in social media: an evaluation. In: ICWSM, Cambridge

Lee K, Tamilarasan P, Caverlee J (2013b) Crowdturfers, campaigns, and social media: tracking and revealing crowdsourced manipulation of social media. In: 7th international AAAI conference on weblogs and social media (ICWSM), Cambridge

Marti S, Garcia-Molina H (2006) Taxonomy of trust: categorizing p2p reputation systems. Comput Netw Int J Comput Telecommun Netw 50(4):472–484

Meier P (2012) What was novel about social media use during hurricane sandy? http://irevolution.net/2012/10/31/hurricane-sandy/

Resnick P, Kuwabara K, Zeckhauser R, Friedman E (2000) Reputation systems. Commun ACM 43(12):45–48

Ross J, Irani L, Silberman MS, Zaldivar A, Tomlinson B (2010) Who are the crowdworkers? Shifting demographics in mechanical turk. In: CHI'10 extended abstracts on human factors in computing systems, CHI EA'10, New York, pp 2863–2872. ACM

Tang JC, Cebrian M, Giacobe NA, Kim H-W, Kim T, Wickert DB (2011) Reflecting on the darpa red balloon challenge. Commun ACM 54(4):78–85

Venetis P, Garcia-Molina H (2012) Quality control for comparison microtasks. In: CrowdKDD 2012, Beijing

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

Wired (2010) The chinese online 'water army'. http://www.wired.com/beyond_the_beyond/2010/06/the-chinese-online-water-army/, June 2010

Xia T, Zhang C, Xie J, Li T (2012) Real-time quality control for crowdsourcing relevance evaluation. In: Network infrastructure and digital content (IC-NIDC), Beijing

Zittrain J (2009) Work the new digital sweatshops. Newsweek