Characterizing and Detecting Malicious Crowdsourcing

Tianyi Wang[†], Gang Wang[‡], Xing Li[†], Haitao Zheng[‡], Ben Y. Zhao[‡]
[†]Electronic Engineering, Tsinghua University, Beijing, China
[‡]Department of Computer Science, University of California, Santa Barbara, USA
wty07@mails.tsinghua.edu.cn, {gangw, htzheng, ravenben}@cs.ucsb.edu, xing@cernet.edu.cn

Categories and Subject Descriptors

J.4 [**Computer Application**]: Social and behaviorial sciences

General Terms

Human Factors, Security

Keywords

Malicious crowdsourcing, crowdturfing, user behavior

1. MALICIOUS CROWDSOURCING

Popular Internet services in recent years have shown that remarkable things can be achieved by harnessing the power of the masses. However, crowd-sourcing systems also pose a real challenge to existing security mechanisms deployed to protect Internet services, particularly those tools that identify malicious activity by detecting activities of automated programs such as CAPTCHAs. Thus they would perform poorly or be easily bypassed when attacks are generated by real human workers inside a crowd-sourcing system.

Through our earlier measurements, we have detect a rapidly growing industry of malicious crowd-sourcing services in multiple countries, including China, India, and the US [4]. In these *crowdturfing* sites, entities such as PR firms, companies, or individuals submit campaigns to the site. Campaigns are usually highly questionable tasks that violate acceptable user practices, such as spreading false rumors via fake tweets, or creating, cloning, or maintaining fake user accounts. Once uploaded, the site divides each campaign into a large number of small tasks (per tweet, or per account requested in the campaign). Tasks are then offered to a large population of members, who are crowd-sourcing workers willing to break the rules for a small payment (usually less than $0.25 per task). Note the involvement of human participants distinguish these campaigns from fake or Sybil accounts controlled by scripts [5, 2, 3].

While these sites are rapidly growing around the world, the largest of these services are two websites hosted in China known as ZhuBaJie and SanDaHa. Records of all transactions on these sites are all public, allowing us to mine them to understand the ecosystem, the entities involved, and the campaigns and tasks processed. Our ongoing studies shows these sites are growing at an exponential pace in both campaigns and amount of revenue generated, and their campaigns have targeted a large range of web sites ranging from microblogging services to instant-messaging networks [4].

Increasing Secrecy. Since records on these sites are currently open to their members, web service admins could hypothetically "defend" against crowdturfing campaigns by tracking all tasks to identify and remove their output from their own sites. In the 12 months since our first study was published, however, several crowdturfing sites have taken steps to gradually make tracking their jobs more difficult. For some new jobs, specific details of the task, *e.g.* content templates or target accounts, are only revealed to workers that take on a task. Since worker accounts require association with phone numbers or bank accounts, this makes tracking jobs significantly more difficult and easier to detect. We expect that further steps will be taken in the near future to make jobs completely invisible (and untrackable) to all except verified worker accounts.

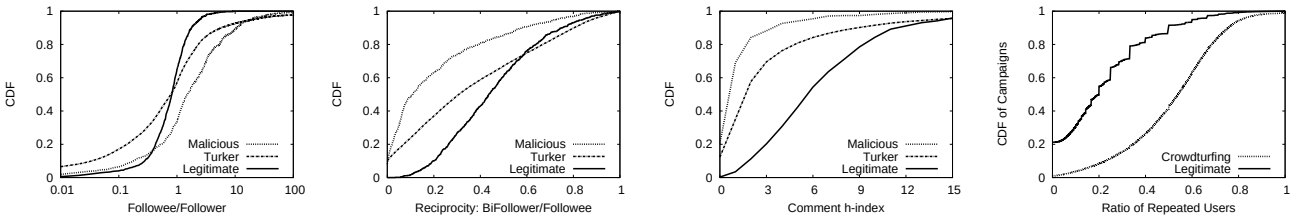
Behavioral Signatures. While tasks on crowdturfing sites are still semi-public, we seek to design and evaluate real-time systems to detect crowdturfing behavior. Ideally, these systems would not rely on information gathering from crowdturfing sites, but would instead perform detection solely by identifying the data output of crowdturfing tasks.

Intuitively, output from crowdturfing tasks are likely to display specific patterns that distinguish them from "organically" generated content. Such differences could be specific to the worker accounts used to perform crowdturfing, *e.g.* their behavior over time, or in the content itself, *i.e.* bursts of content generation when tasks are first posted. Therefore, understanding the characteristic "signatures" of crowdturfing activity requires comparing their output to that of normal content. Fortunately, these sites have yet to fully implement privacy measures, and for now, we still have direct access to records of crowdturfing campaigns and tasks, which allow us to directly track their output and gather "ground truth" to compare against organic content.

Our Methodology. In this work, we seek to gain a deep understanding of the identifiable characteristics (and signatures) found in the output of crowdturfing campaigns and tasks. We note that many of these signatures are likely to be application-specific. In our first step, we limit our scope to campaigns that target microblogging platforms, *e.g.* Twitter and Sina Weibo.

Our approach includes two phases of work. First, we gather a large volume of "ground truth" content that has been identified and matched with specific tasks on known crowdturfing sites. We also gather "organic" content generated by normal users. Second, we compare and contrast these datasets with respect to both the

SIGCOMM'13, August 12–16, 2013, Hong Kong, China.
ACM 978-1-4503-2056-6/13/08.



(a) Ratio of Followee/Follower (b) Reciprocity (c) Comment h-index (d) Repeated Users

Figure 1: Comparing the social profile and activity patterns of crowdturf workers and normal accounts.

users (account profiles and temporal activity patterns of workers vs. those of normal users) and output content (analysis of embedded URLs and keywords, bursty patterns in content generation, and popularity or retweet patterns).

These ground truth datasets give us a valuable platform for designing and experimentally evaluating crowdturf detection systems. Our end goal is to develop and continually fine-tune detectors by testing them against new crowdturfing campaigns as they arrive. When ready, we will release these detectors and offer them to potential targets such as microblogging sites (Weibo, Twitter), social networks (Facebook, Renren) and chat forums such as QQ.

2. MEASURED DATASETS

Since our first work on crowdturfing sites in early 2012, we have continued to actively gather data from these sites and the output of their campaigns on a number of platforms. Our measurements have focused on the two largest crowdturfing sites we know, ZhuBaJie (ZBJ) and SanDaHa (SDH). We have extracted records of all campaigns to ever appear on these sites, more than 290,000 campaigns in total (over 6 years for ZBJ and 3 years for SDH). On average, each campaign generates between 50 to 100 individual tasks.

Crowdturf Accounts on Weibo. Focusing on the Sina Weibo microblogging platform (essentially China's own Twitter), we have extracted the Weibo account identifiers from transaction records of all workers who have completed crowdturfing tasks. The result is 34,505 Weibo account IDs, of which 5,558 have already been blocked by Sina Weibo. For the remaining 28,947 Weibo accounts, we have downloaded full user profiles, following relationships and the latest 2,000 tweets from each account.

Crowdturf Campaigns. We extract 20,416 campaigns that target Weibo, which generated a total of 26,896 tasks. Campaigns generally ask workers to generate tweets, retweet, or post comments on existing tweets (Weibo allows comments on tweets, unlike Twitter). Tweets from 2,081 campaigns have already been deleted by the authors or Weibo. We crawled all tweets, retweets and comments of the remaining 18,335 campaigns. As a basis for comparison, we also crawled (from Nov. 2012 to Jan. 2013) 61.5 million tweets, 118 million comments and 86 million retweets, all from 723K users.

Finally, we obtain several sets of accounts with known properties. After crawling the 723K users, we revisited them and found that Weibo had banned roughly 1000 accounts. We also ask a group of student volunteers to manually examine and identify roughly 1000 "legitimate" accounts randomly chosen from the set.

3. SOME INITIAL RESULTS

While our experiments are early and ongoing, they are generally grouped into the following categories. First, we study the profile and activity patterns of Weibo accounts who are active participants in crowdturfing tasks. This includes social connectivity and delays between user-generated events. Second, we analyze crowdturf output for a) presence of similar URLs and identifiable keywords [1], b) burstiness of arrival times which may synchronize with arrival of new campaigns, and c) follow-on activity generated by the content, *e.g.* retweets and comments.

For space constraints, we only show a small sample of our results in Figure 1. When we compare our sets of users (banned/malicious, legitimate users and turker/workers), we find that turkers (crowdturfers) tend to straddle the line between malicious and normal users in terms of their social structure, *i.e.* ratio of the users they follow to the number of followers they have, their reciprocity, *i.e.* the portion of users they follow who follow them back, and their h-index values. A user with h-index of h has at least h tweets each with h comments. Not surprisingly, this result says that many crowdturf workers likely use their accounts normally at least part of the time, producing an account profile that is more similar to legitimate users and is thus harder to identify.

We also study the patterns of crowdturfing and legitimate campaigns. We show the ratio of repeated users in the campaigns. There is a clear trend that crowdturfing campaigns have a higher ratio of repeated users, *i.e.*, users who retweet on this campaign multiple times. We will work on more campaign patterns in future works and develop detectors based on these features.

4. ACKNOWLEDGEMENT

This project is supported by Tsinghua University Initiative Scientific Research Program [No.20111081023, No.20111081010] and National High-tech R&D Program of China (863)[2012AA011004]. This project is also partially supported by NSF grant CNS-1224100 and DARPA BAA 12-01.

5. REFERENCES

[1] GAO, H., HU, J., WILSON, C., LI, Z., CHEN, Y., AND ZHAO, B. Y. Detecting and characterizing social spam campaigns. In *Proc. of IMC* (2010).

[2] WANG, G., KONOLIGE, T., WILSON, C., WANG, X., ZHENG, H., AND ZHAO, B. Y. You are how you click: Clickstream analysis for sybil detection. In *Proc. of USENIX Security* (Washington, D.C., August 2013).

[3] WANG, G., MOHANLAL, M., WILSON, C., WANG, X., METZGER, M., ZHENG, H., AND ZHAO, B. Y. Social turing tests: Crowdsourcing sybil detection. In *Proc. of NDSS* (San Diego, CA, February 2013).

[4] WANG, G., WILSON, C., ZHAO, X., ZHU, Y., MOHANLAL, M., ZHENG, H., AND ZHAO, B. Y. Serf and turf: Crowdturfing for fun and profit. In *Proc. of WWW* (2012).

[5] YANG, Z., WILSON, C., WANG, X., GAO, T., ZHAO, B. Y., AND DAI, Y. Uncovering social network sybils in the wild. In *Proc. of IMC* (2011).