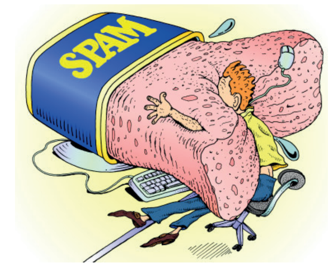
Understanding and Defending Against Malicious Crowdsourcing

Ben Y. Zhao (PI), Haitao Zheng (CoPI), Gang Wang (PhD student), University of California at Santa Barbara

SANDLab
http://sandlab.cs.ucsb.edu
ravenben@cs.ucsb.edu

1. Malicious Crowdsourcing

New Threat: Malicious Crowdsourcing = Crowdturfing



+ Hire a large group of real Internet users for malicious attacks
+ Fake reviews, rumors, targerted spam
+ Most existing defenses failed against real users (*e.g.*, CAPTCHA)

Crowdturfing Sites

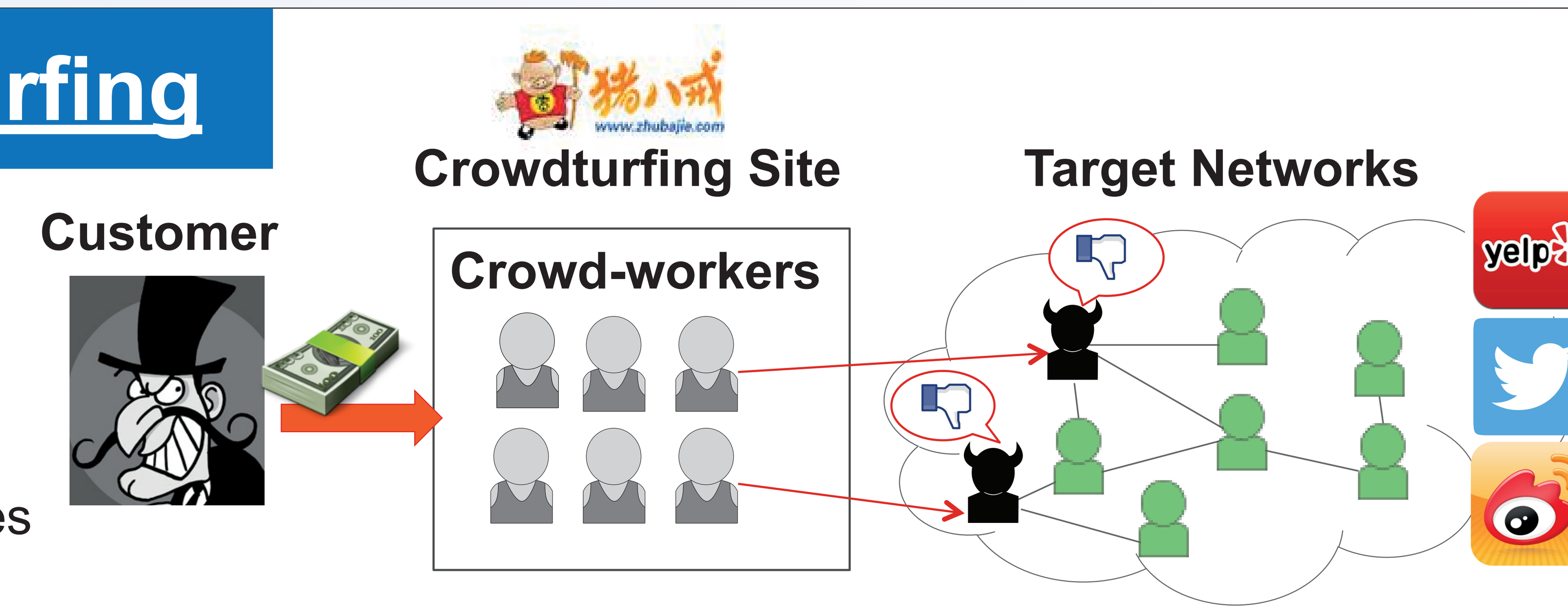
+ Web services that recruit Internet users as workers (spam for $)
+ Connect workers to customers who want to run malicious campaigns

Research Questions

+ How does crowdturfing work? [1]
+ What's the scale, economics and impact of crowturfing campaigns? [1]
+ How to defend against crowdturfing? [2]

2. Understanding Crowdturfing

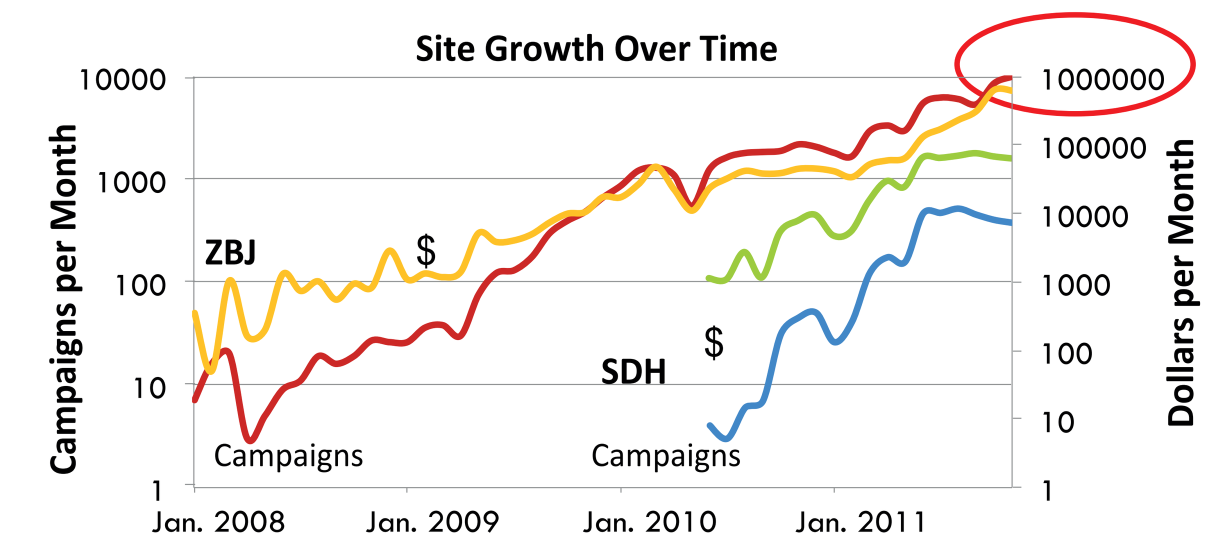
Key Players



+ **Customers:** pay to run a campaign
+ **Workers:** real users, spam for $
+ **Target Networks:** social networks, revew sites

Scale and Revenue

+ Measurements of two largest crowdturfing sites (in China)
  - ZBJ (zhubajie.com), five years
  - SDH (sandaha.com), two yeras
+ 18.5M tasks, 79K campaigns, 180K workers
+ Millions dollars of revenue per month



Crowdturfing around the World

ZBJ, SDH Fiverr, Freelancer, MinuteWorkers, Myeasytasks, Microworkers, Shorttasks Paisalive

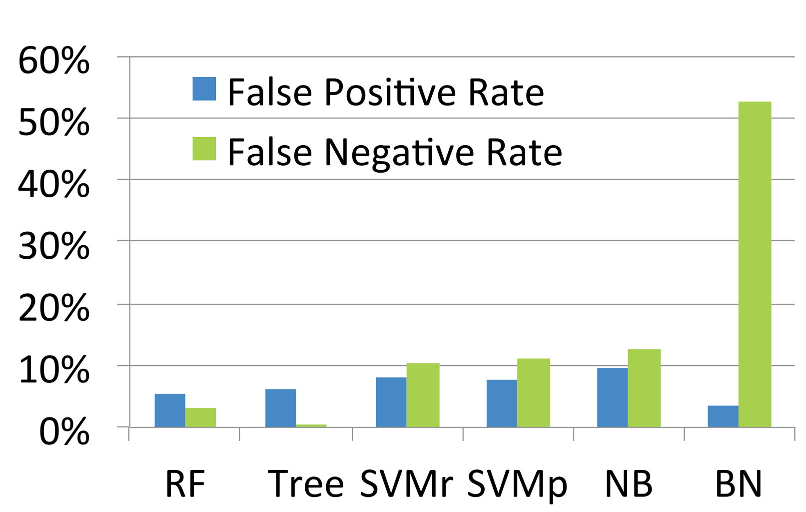
3. Defense: Machine Learning Classifiers

Machine Learning (ML) vs. Crowdturfing

+ Simple method does not work on real users (*e.g.,* CAPTCHA, rate limit)
+ **Machine learning:** more sophiscaed modeling on user behaviors
+ Perfect context to study adversarial machine learning
  - **Human workers** are adaptive to evade classifiers
  - **Crowdturf admins** can temper with training data by chaning worker behaviors

How Effective is ML-based Detecor?

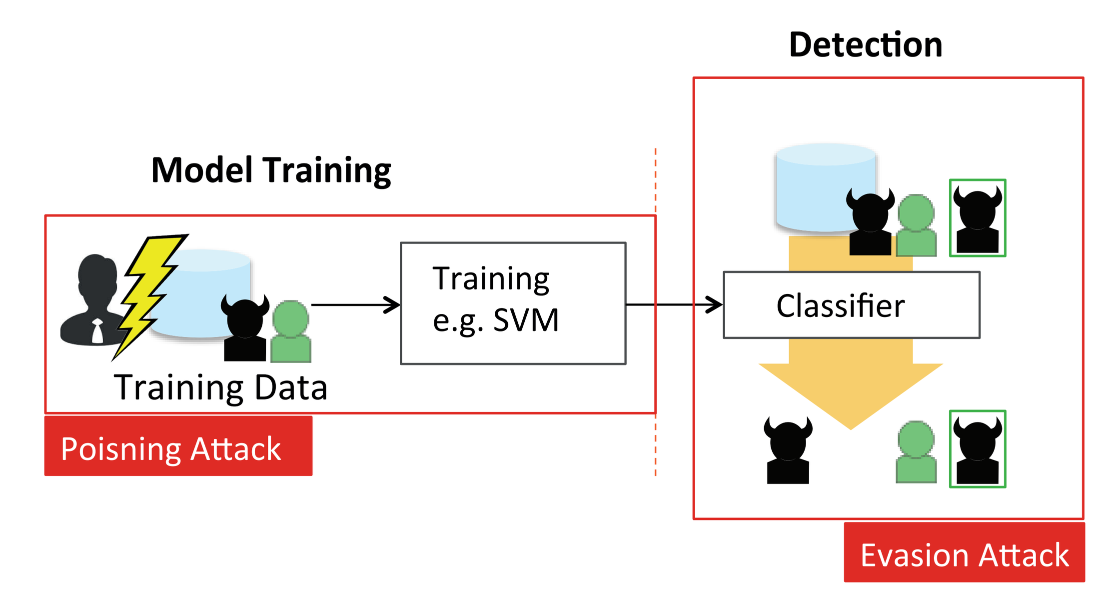
+ **Groundtruth:** 28K workers in crowdturfing campaigns on *Weibo* (Chinenes Twitter)
+ **Baseline users:** 371K Weibo user accounts
+ 30 behavioral features
+ **Classiiers:** Random Forest, Decision Tree, SVM, Naive Bayes, Bayesian Network



+ Random Forest is the most accurate (95% accuracy)
+ 99% accuracy on professional workers (>100 tasks)
+ How **robust** are those classifiers?

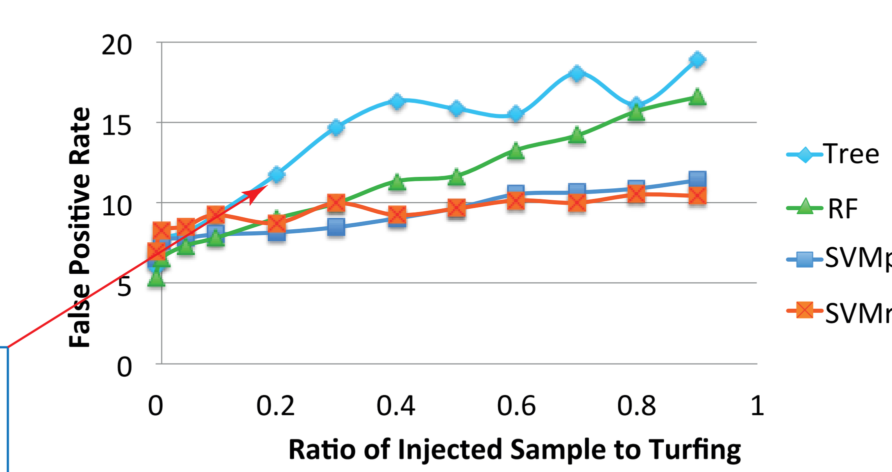
Adversarial Machine Learning

+ **Evasion attack:** individual workers change behaviors to evade the detection
  - Impact: single feature-change saves 95% of workers
+ **Poisoning attack:** site admins tamper with training data to mislead classifier training



Example: Poisoning Attack

+ Inject mislabeled samples to training data ➔ wrong classifier
  e.g., inject benign accounts as "workers" in training data
+ Uniformly change workers behavior by enforcing task policies
  ➔ hard to train an accurate classifier



More accurate classifiers can be more vulnerable

Summary

+ Machine learning classifiers are effective against current crowd-workers
+ Classifiers are highly vulnerable to adversarial attacks. Future works will focus on improving the robustiness of ML-classifiers

[1] G. WANG, T. WANG, H. ZHENG, B. ZHAO. Man vs. machine: practical adversarial detection of malicious crowdsourcing workers. In *Proc. of Usenix Security* (2014)

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