Towards Robust Crowd Computations

Alan Mislove
Northeastern University

Motivation

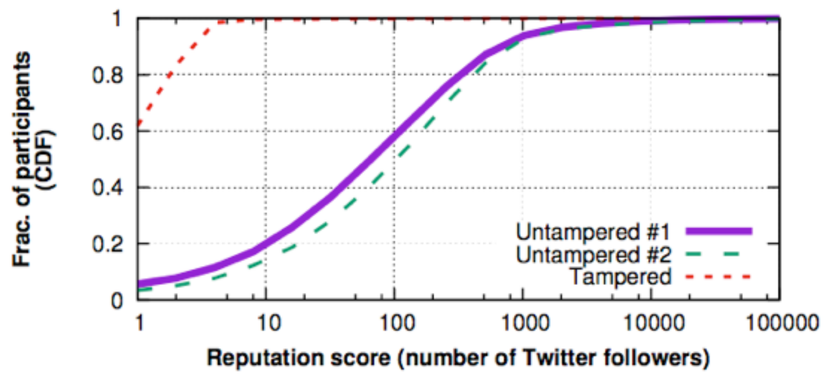
Sybil attacks on crow computation systems: An attacker create fake identities and manipulate the aggregate opinion of the crowd (e.g., rating of business on Yelp, # of followers in Twitter)

Existing approach to detect individual Sybil identities has a **fundamental limitation**: Adaptive attacker can create hard-to-detect Sybil identities

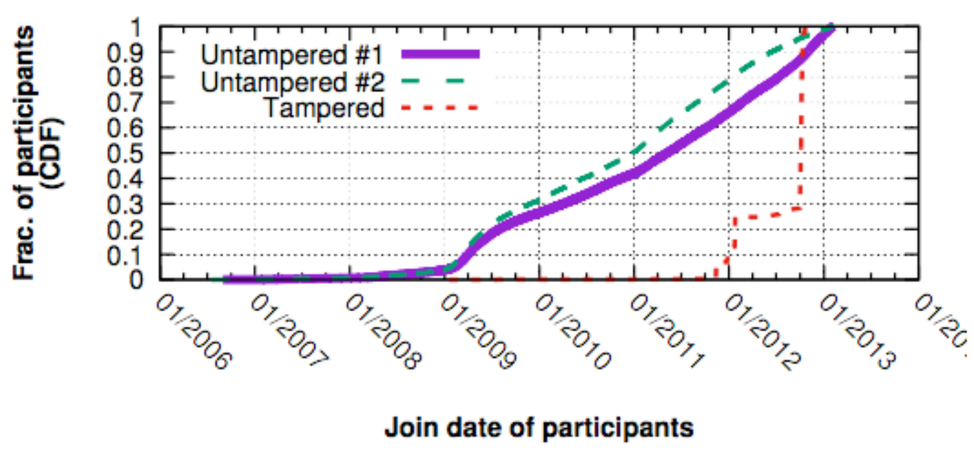
We developed **Stamper**, which detects a tampered computation robustly

Key Insights

1. Sybil identities as a group, **tend to skew** towards low reputation scores (e.g., number of Twitter followers)



2. Sybil identities can't change the **past history** (e.g., join date)



Focus on the **group** of sybil identities and **unforgeable timestamps** of activities

Stamper Algorithm

Goal: Detect whether a given large crowd computation was tempered by Sybil identities or not

Algorithm:
(1) Compare distributions of reputation scores (e.g., # of friends in a crowd or join dates) using Kullback-Leibler (KL) divergence (i.e., all vs. target sets)

$$\mathrm{KLD}(P,Q) = \sum_{i=1}^{r} \left(\log\left(\frac{P(i)}{Q(i)}\right) P(i) + \log\left(\frac{Q(i)}{P(i)}\right) Q(i) \right)$$

(2) If KLD(P, Q) < ε, it means that two distributions are similar (i.e., target sets are NOT tampered)
(3) If KLD(P, Q) > ε, it means that two distributions are very different (i.e., target sets are tampered)

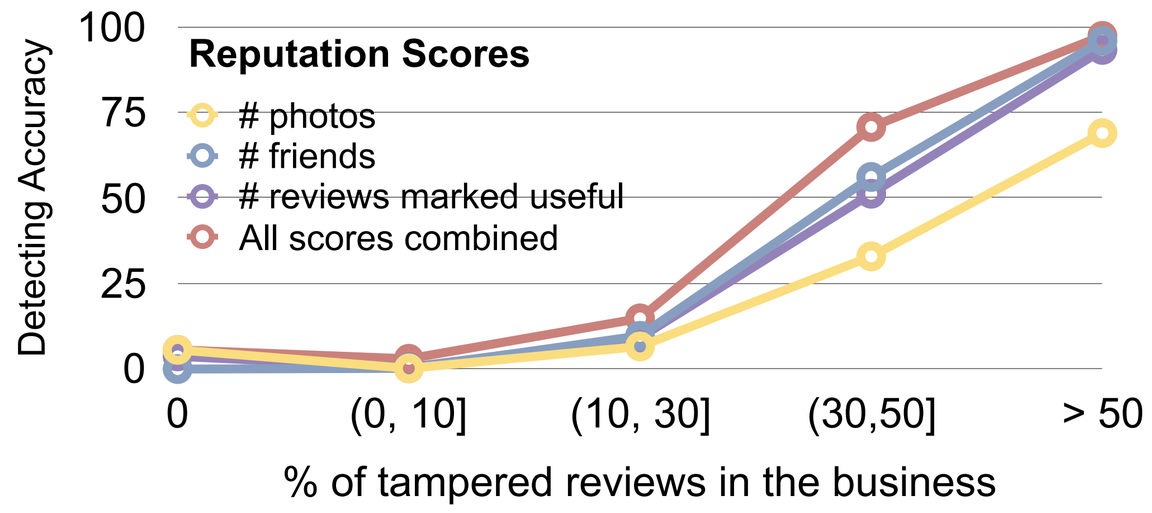
Evaluation

Case 1: Yelp Review Tampering on Business

Goal: Find the business with highly tampered reviews

Data	Numbers
# of all business	3,579
# of tampered reviews	195,825
# of business not having tampered reviews*	54

*Each business has different portion of tampered reviews



Stamper can detect most of the highly tampered (> 50%) business while having low false positive rate (only 3%)

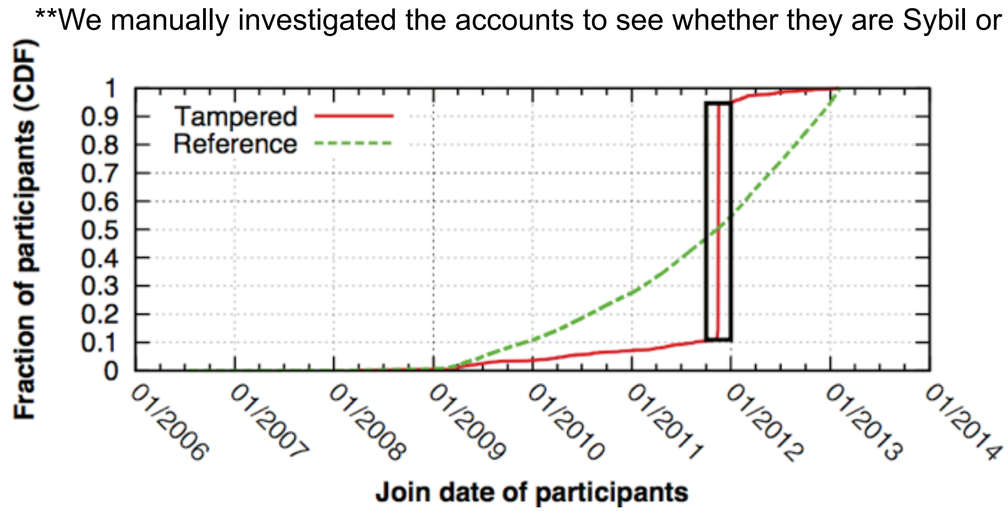
Case 2: Twitter Follower Tampering

Goal: Find the Twitter accounts having tampered followers

Data	Numbers
# of users	69,409
# of users not having tampered followers*	30,000
# of found Sybil accounts from **Stamper** using *join dates* distribution**	620

*These accounts are verified from Twitter

**We manually investigated the accounts to see whether they are Sybil or not



Unforgeable timestamp is an effective measure to detect tampering

Conclusion

- Stamper can detect tampered computation (by Sybil)
- Key insight is focus on (1) large statistical samples (groups) of Sybil and (2) unforgeable timestamp information
- Stamper can raise the bar for defense against adaptive attackers and detect regardless of attacker strategy

Interested in meeting the PIs? Attach post-it note below!

National Science Foundation
WHERE DISCOVERIES BEGIN

NSF SaTC Principal Investigator Meeting
Jan 9-11, 2017
Alexandria, VA