# Understanding, Predicting, & Influencing Human Decision Making

**Krishna P. Gummadi**

**Max Planck Institute for Software Systems**

# Outline for my lectures

- Lecture 1:
  - Overview of (my) Social Computing Research
- Lecture 2:
  - On the Temporality of Trust and Privacy
- Lecture 3:
  - On Biases in Search & Recommendations in Crowdsourcing Systems

# Time is a fascinating dimension

- We obsess and struggle to manage time well
  - Path 1: Effort_1(t), Payoff_1(t)
  - Path 2: Effort_2(t), Payoff_2(t)

- The challenge lies in estimating efforts and payoffs
  - The functions are personalized!
  - The functions are biased by your social circle!!
  - The functions don't compose linearly!!!
  - The function estimations vary with time!!!!

# When is some effort worth funding?

- What is its corresponding Payoff(t) function?
  - If Payoff(t) is a head-heavy distribution and certain:
    - It is called applied research
    - It will be funded very competitively by industry
  - If Payoff(t) is a long-tail and uncertain:
    - It is called foundational research
    - It will have to be funded by academic researchers
    - Based on criteria like elegance, novelty, beauty, depth, truth

# Finding new problems over time

- Time is a great dimension for thought experiments
  - Frequently, helps discover new problems
- Example Problem: Fairness in decision making
  - Do our assessments of fairness have a time dimension?
  - Is the fairness of a decision making system time-invariant?
- Can we design incrementally fair decision systems?

# This talk

Explore temporal dimensions of

- Trust in social computing systems
- Privacy in social computing systems

# Examples of social computing systems

- Social networking sites: Facebook, Goolge+
- Blogging sites: Twitter, LiveJournal
- Content-sharing sites: YouTube, Flickr
- Social bookmarking sites: Delicious, Reddit
- Crowd-sourced opinions: Yelp, eBay seller ratings
- Peer-production sites: Wikipedia, AMT

- Distributed systems of people

# The Achilles Heel of Social Computing Systems

# Trust in identity infrastructures

Most platforms use a weak identity infrastructure

Weak identity infrastructure:

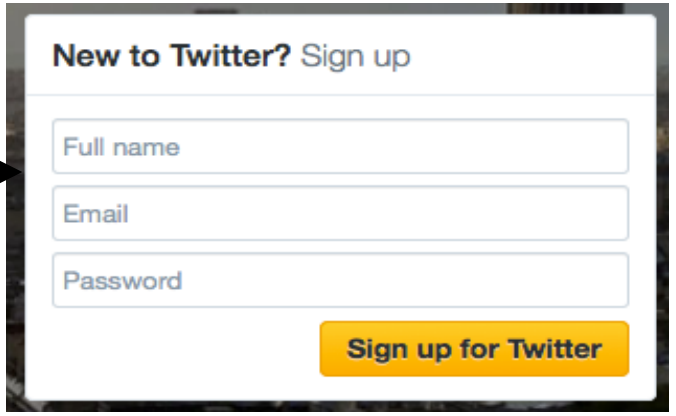
No verification by trusted authorities required. Fill up a simple profile to create account

Pros:

- Provides some level of anonymity
- Low entry barrier

Cons:

- Lack accountability
- Vulnerable to fake (Sybil) id attacks



# Sybil attacks: Attacks using fake identities

- Fundamental problem in systems with weak user ids
- Numerous real-world examples:
  - Facebook: Fake likes and ad-clicks for businesses and celebrities
  - Twitter: Fake followers and tweet popularity manipulation
  - YouTube, Reddit: Content owners manipulate popularity
  - Yelp: Restaurants buy fake reviews
  - AMT, freelancer: Offer Sybil identities to hire

# Sybil attacks are a growing menace

There is an incentive to manipulate popularity of ids and information

**Harvard Study: Yelp Drives Demand for Independent Restaurants**

"[A] one-star increase in Yelp rating leads to a 5-9% increase in revenue...

**Syncapse: Each Facebook Like Is Worth $174 To Brands**

THE WALL STREET JOURNAL.
WSJ.com

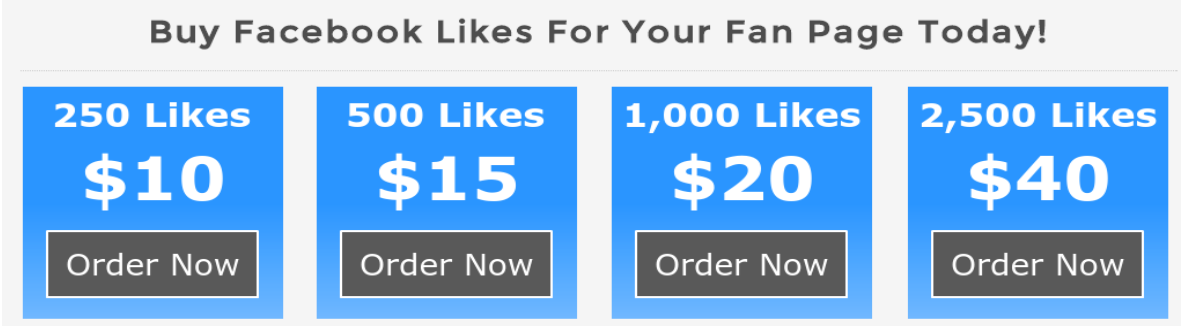
March 28, 2014, 2:43 PM ET

**What's More Valuable: A Stolen Twitter Account or a Stolen Credit Card?**

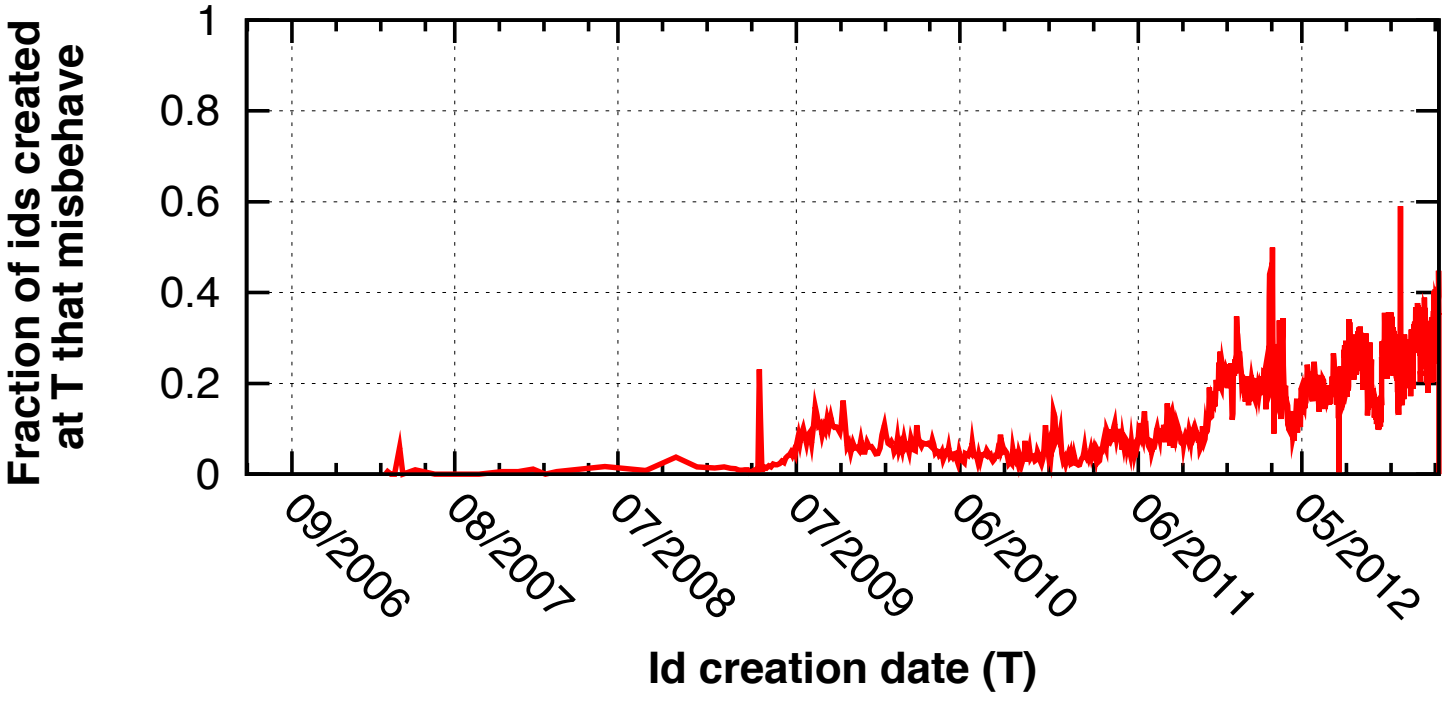
ByElana Zak

# The emergence of Abuse-As-A-Service

## After Sting Operation, Yelp Outs 8 Businesses That It Caught Trying To Buy Reviews

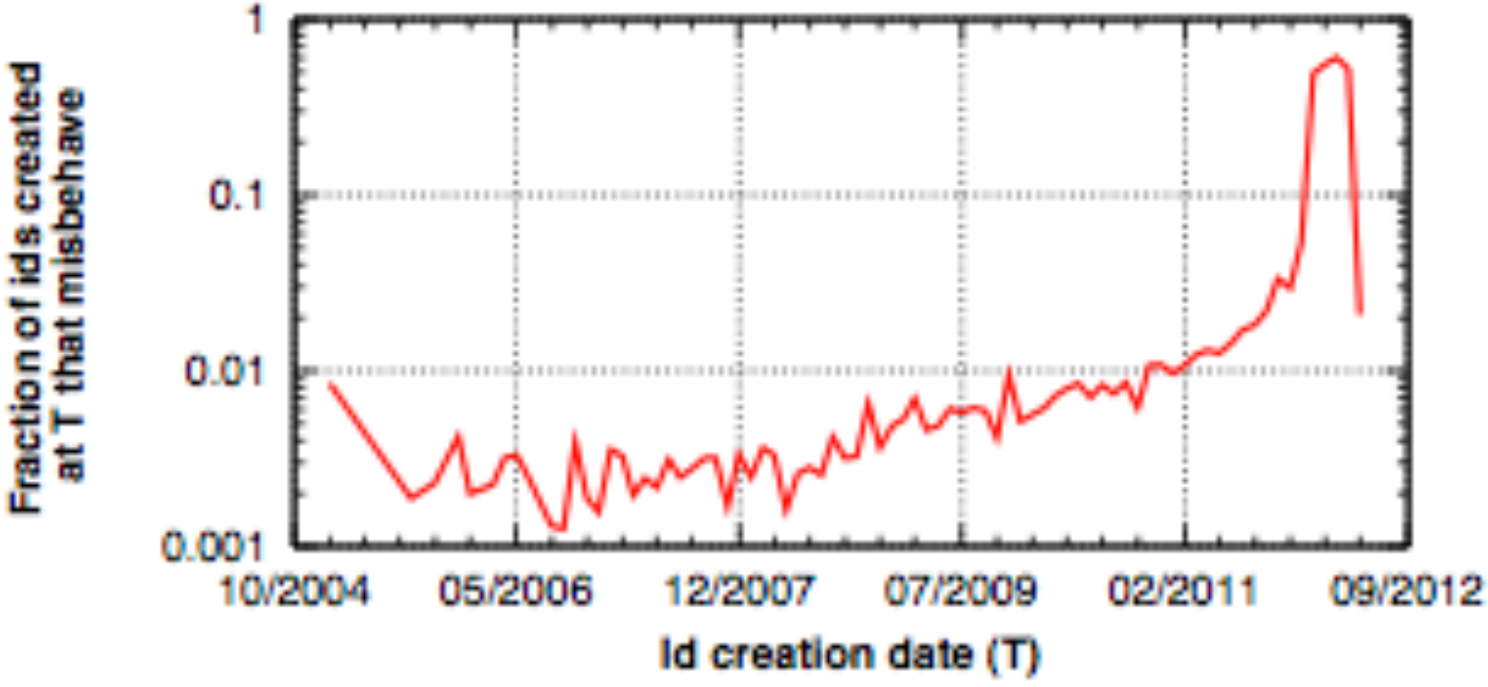



# Sybil identities are a growing menace




- 40% of all newly created Twitter ids are fake!

# Sybil identities are a growing menace




- 50% of all newly created Yelp ids are fake!

# The Strength of Weak Identities

# Strength of a weak identity

- Effort needed to forge the weak identity
- Weak ids come with zero external references
- Strength is the effort needed to forge ids' activities
  - And thereby, the ids' reputation
- Idea: Could we measure ids' strength by their blackmarket prices?

| Domain | Price range per account ($) | Median price per account ($) |
|---|---|---|
| Hotmail | 0.003 to 0.45 | 0.013 |
| Yahoo | 0.01 to 0.375 | 0.038 |
| Twitter | 0.010 to 1 | 0.093 |
| Pinterest | 0.05 to 0.5 | 0.103 |
| Google | 0.033 to 0.67 | 0.145 |
| LinkedIn | 0.05 to 0.5 | 0.250 |
| Facebook | 0.10 to 2.50 | 0.515 |

**Table 1: Black market prices of Sybils (without any particular reputation).**

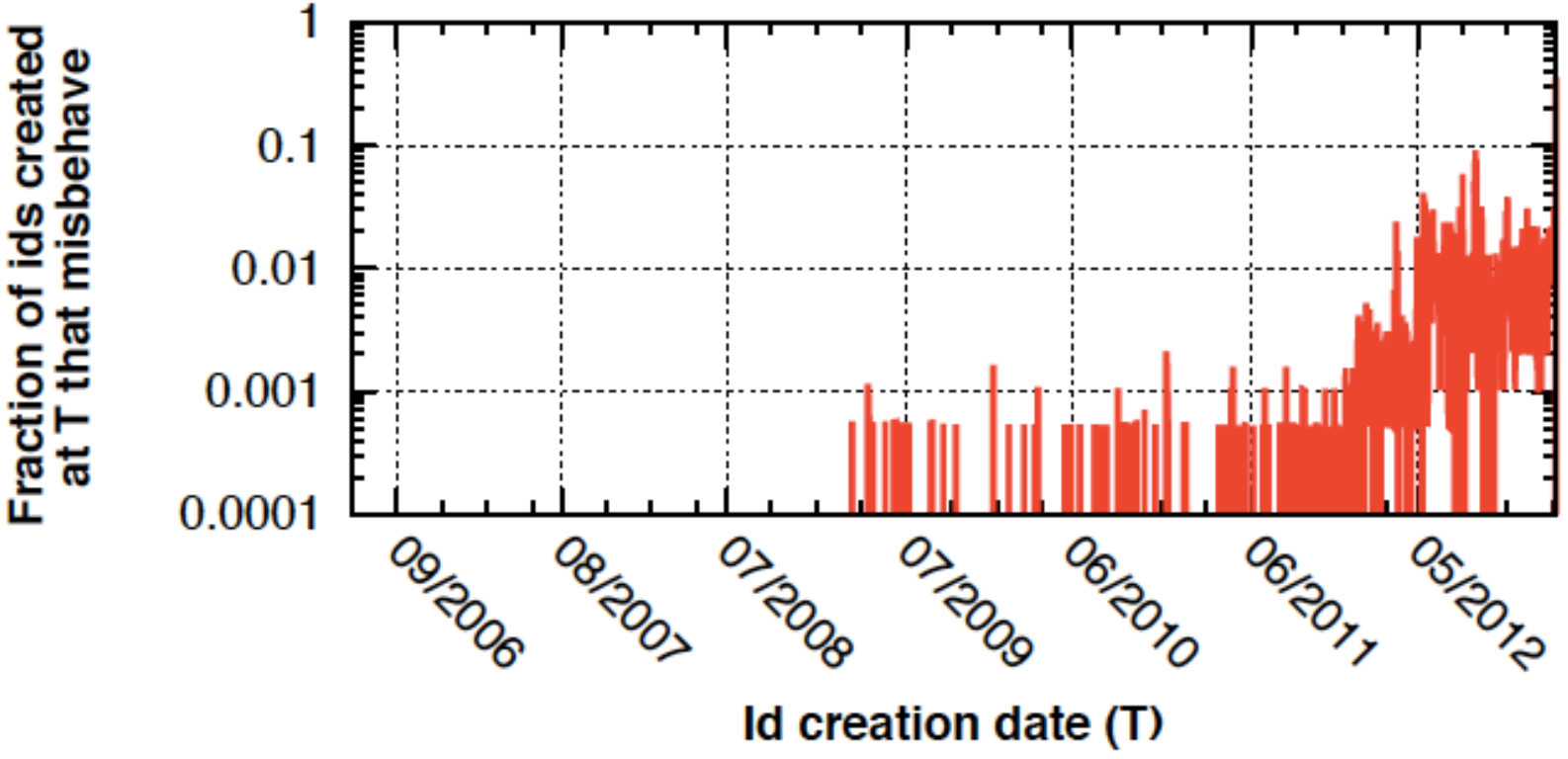
| Domain | Reputation Measure | Price range per account($) |
|---|---|---|
| Twitter | aged 2.5 years | 0.25 |
| Twitter | aged 4 years | 1 |
| Twitter | 100+ followers | 0.5 |
| Twitter | 300+ followers | 1 |
| Twitter | 200+ real/active followers | 5 |
| Facebook | aged 1.5 years | 5 to 6 |
| Facebook | aged 4 years | 15 to 16 |
| Facebook | 1000 real/active friends | 30 |
| Facebook | 5000 real/active friends | 150 |

**Table 2: Black market prices of Sybil identities with different levels of reputation.**

# Key observation

- Attackers cannot tamper timestamps of activities
  - E.g., join dates, id creation timestamps
- Older ids are less likely to be fake than newer ids
  - Attackers do not target till sites reach critical mass
  - Over time, older ids are more curated than newer ids
    - Spam filters had more time to check older ids

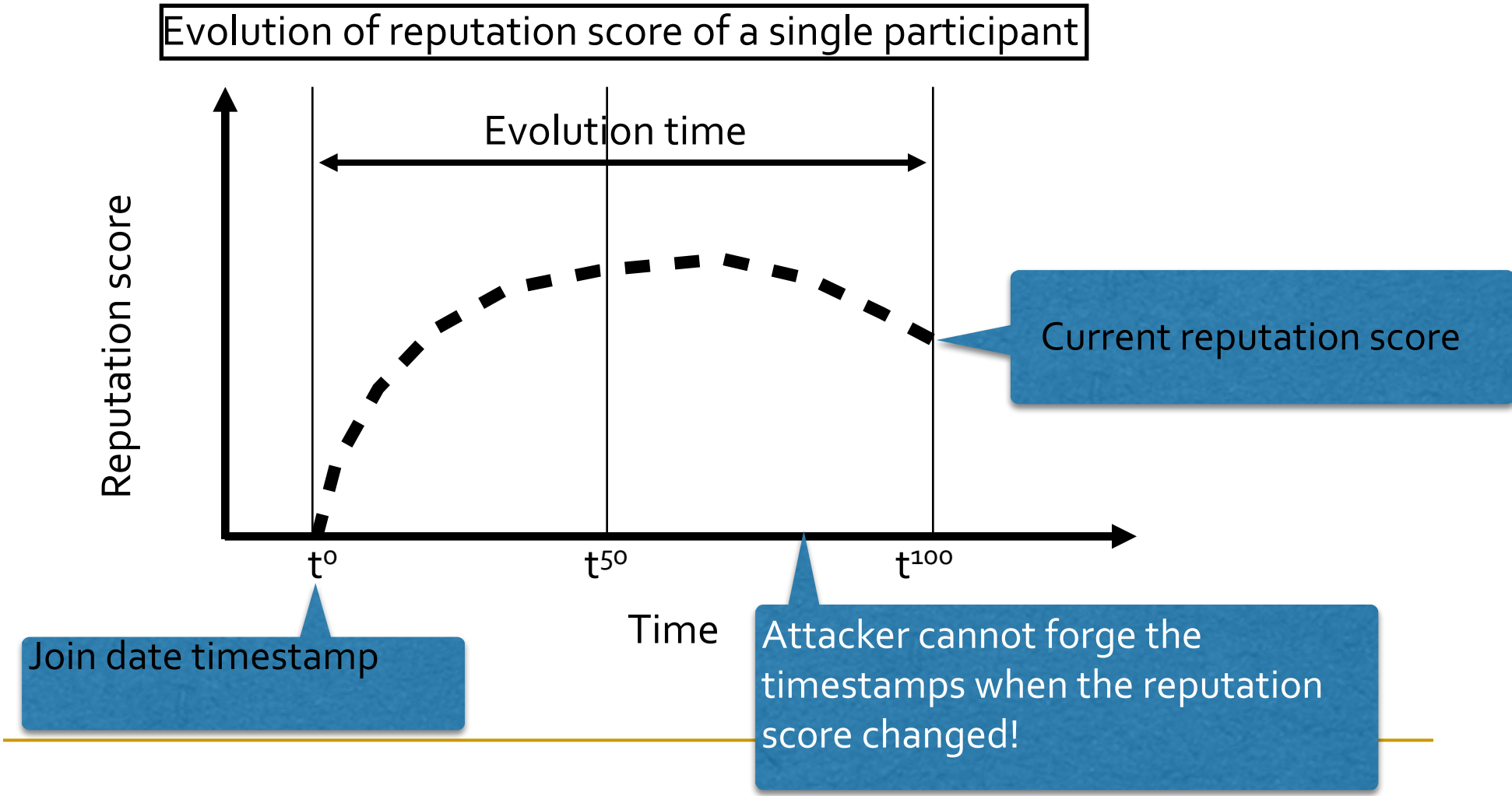
# Most active fakes are new ids




Older ids are less likely to be fake than newer ids

# Assessing strength of weak identities

- Leverage the temporal evolution of reputation scores



# Trustworthiness of Weak Identities

# Trustworthiness of an identity

- Probability that its activities are in compliance with the online site's ToS
- How to assess trustworthiness?
  - Ability to hold the user behind the identity accountable
    - Via non-anonymous strong ids
  - Economic incentives vs. costs for the attack
    - Strength of weak id determines attacker costs
  - Leverage social behavioral stereotypes

# Traditional Sybil defense approaches

- Catch & suspend ids with bad activities
  - By checking for spam content in posts
  - Can't catch manipulation of genuine content's popularity
- Profile identities to detect suspicious-looking ids
  - Before they even commit fraudulent activities
- Analyze info available about individual ids, such as
  - Demographic and activity-related info
  - Social network links

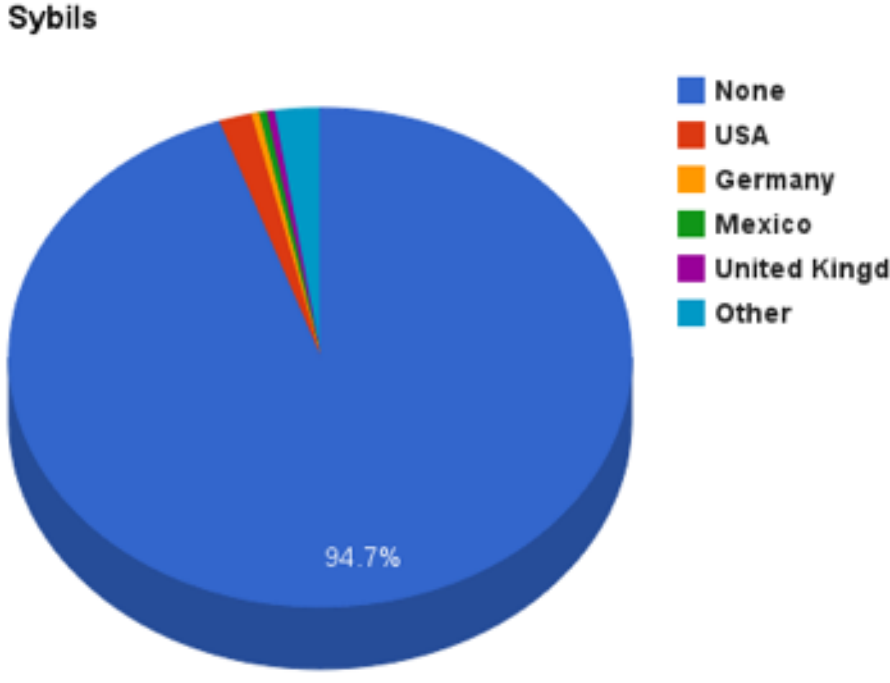
# Lots of recent work

- Gather a *ground-truth* set of Sybil and non-Sybil ids
  - Social turing tests: *Human verification* of accounts to determine Sybils *[NSDI '10, NDSS '13]*
  - Automatically flagging *anomalous (rare)* user behaviors *[Usenix Sec. '14]*

- Train ML classifiers to distinguish between them *[CEAS '10]*
  - Classifiers trained to flag ids with similar profile features
  - Like humans, they look for features that arise suspicion
    - Does it have a profile photo? Does it have friends who look real? Do the posts look real?

# Key idea behind id profiling

- For many profile attributes, the values assumed by Sybils & non-Sybils tend to be different

# Key idea behind id profiling

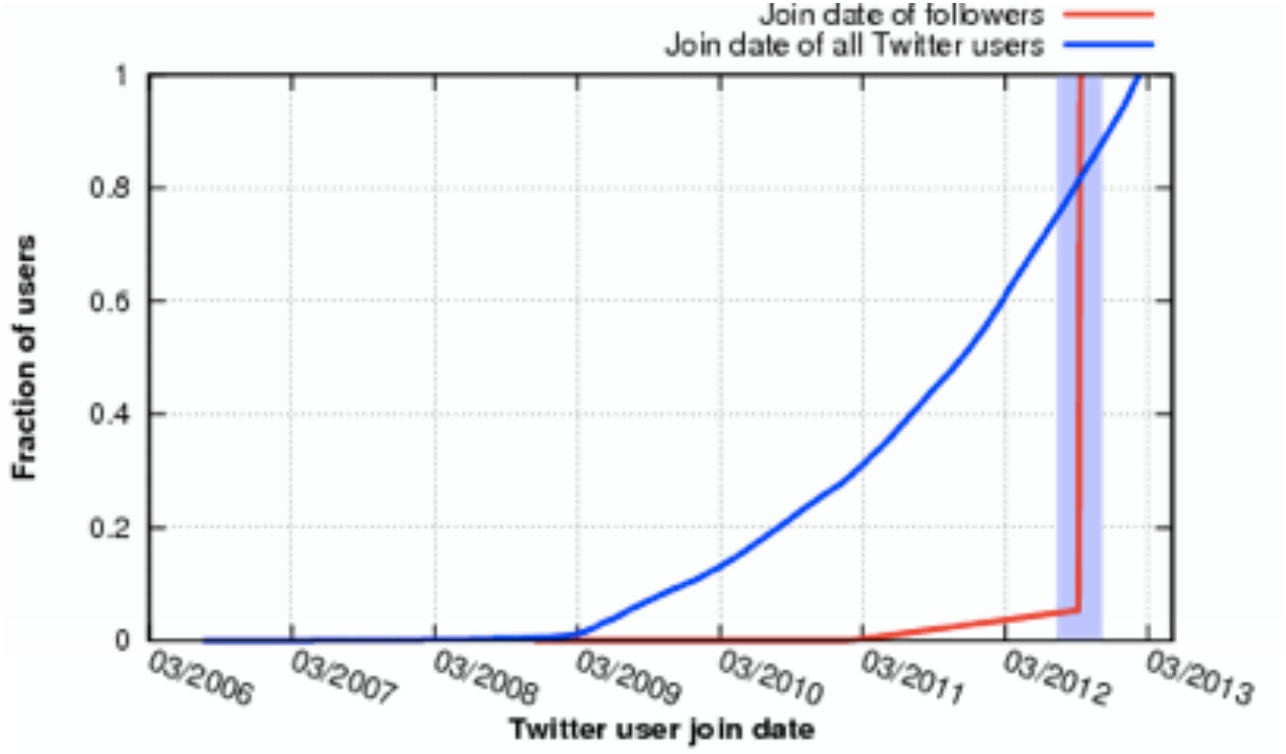
- For many profile attributes, the values assumed by Sybils & non-Sybils tend to be different
  - Location field is not set for >90% of Sybils, but <40% of non-Sybils
  - Lots of Sybils have low follower-to-following ratio
  - A much smaller fraction of Sybils have more than 100,000 followers

# Limitations of profiling identities

- Potential discrimination against good users
  - With rare behaviors that are flagged as anomalous
  - With profile attributes that match those of Sybils

- Sets up a rat-race with attackers
  - Sybils can avoid detection by assuming *likely* attribute values of good nodes
    - Sybils can set location attributes, lower follower to following ratios
  - Or, by attacking with new ids with no prior activity history

# Attacks with newly created Sybils



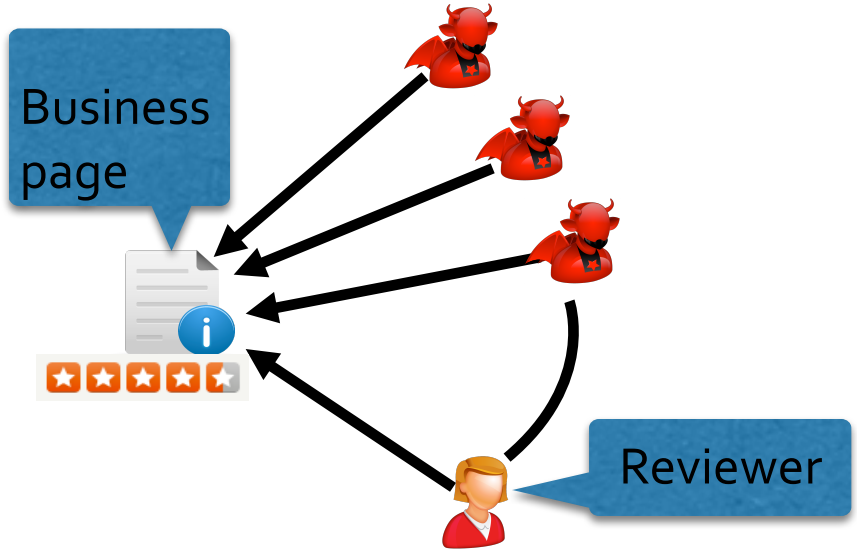

- All our bought fake followers were newly created!
- Existing spam defenses cannot block them

# Robust Tamper Detection in Crowd Computations

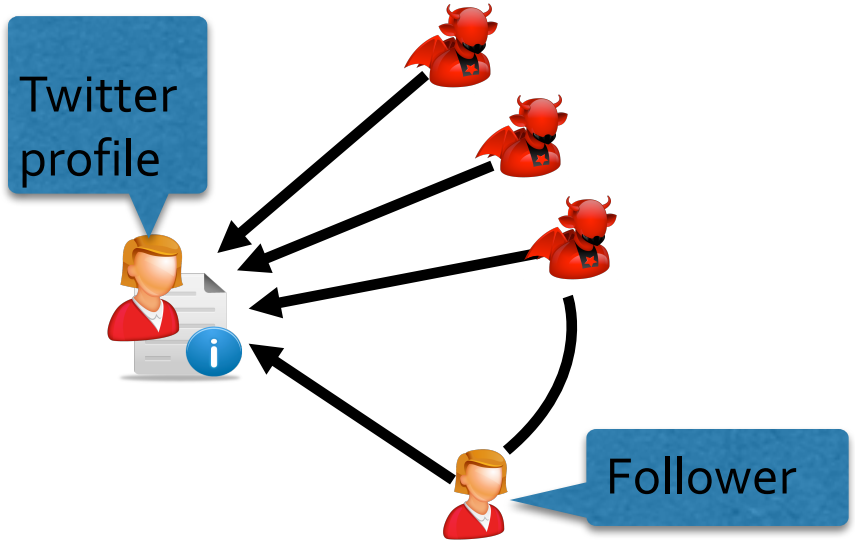
# Is a crowd computation tampered?

Does a large computation involve a sizeable fraction of Sybil participants?

Business with tampered rating




User with tampered follower count



# Are the following problems equivalent?

1. Detect whether a crowd computation is tampered
   - Does the computation involve a sizeable fraction of Sybil participants?

2. Detect whether an identity is Sybil

# Are the following problems equivalent?

1. Detect whether a crowd computation is tampered
   - Does the computation involve a sizeable fraction of Sybil participants?

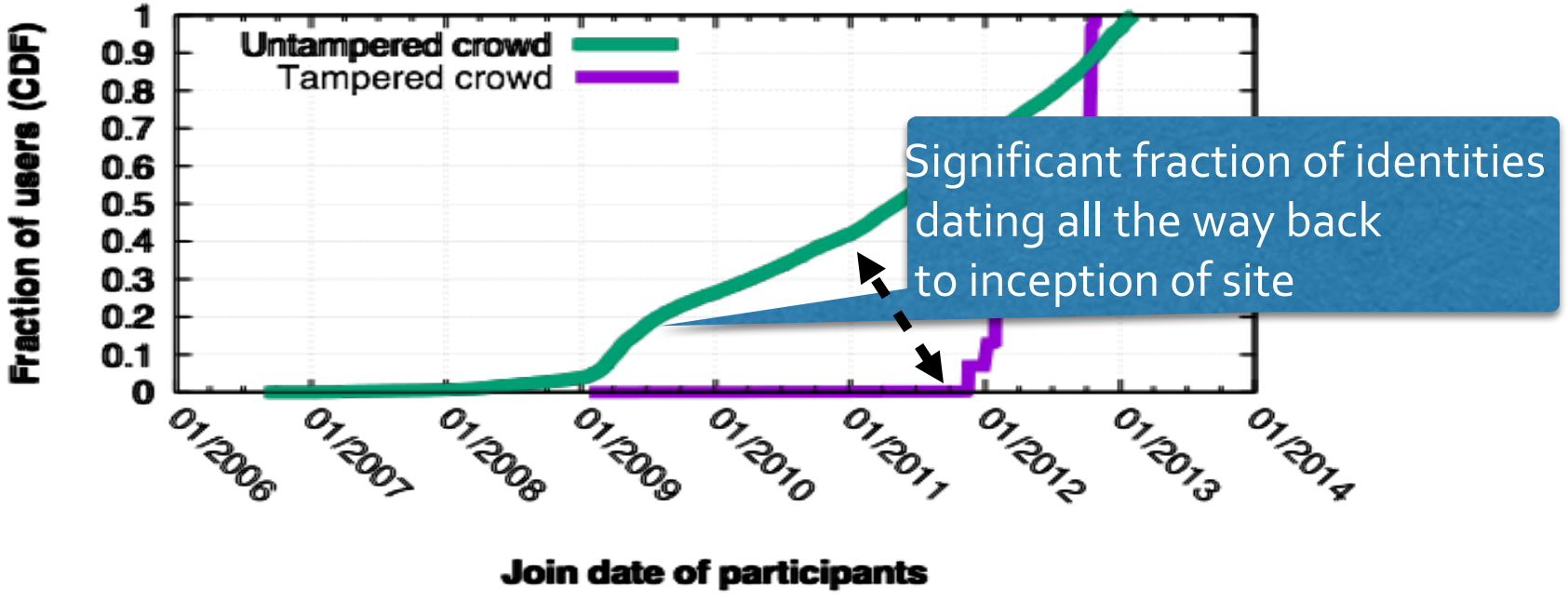
2. Detect whether an identity is Sybil

Our Stamper project: NO!

Claim: We can robustly detect tampered computations even when we cannot detect fake ids

# Stamper: Detecting tampered crowds

- Idea: Analyze join date distributions of participants
  - Entropy of tampered computations tends to be lower
  - More generally, temporal evolution of reputation scores

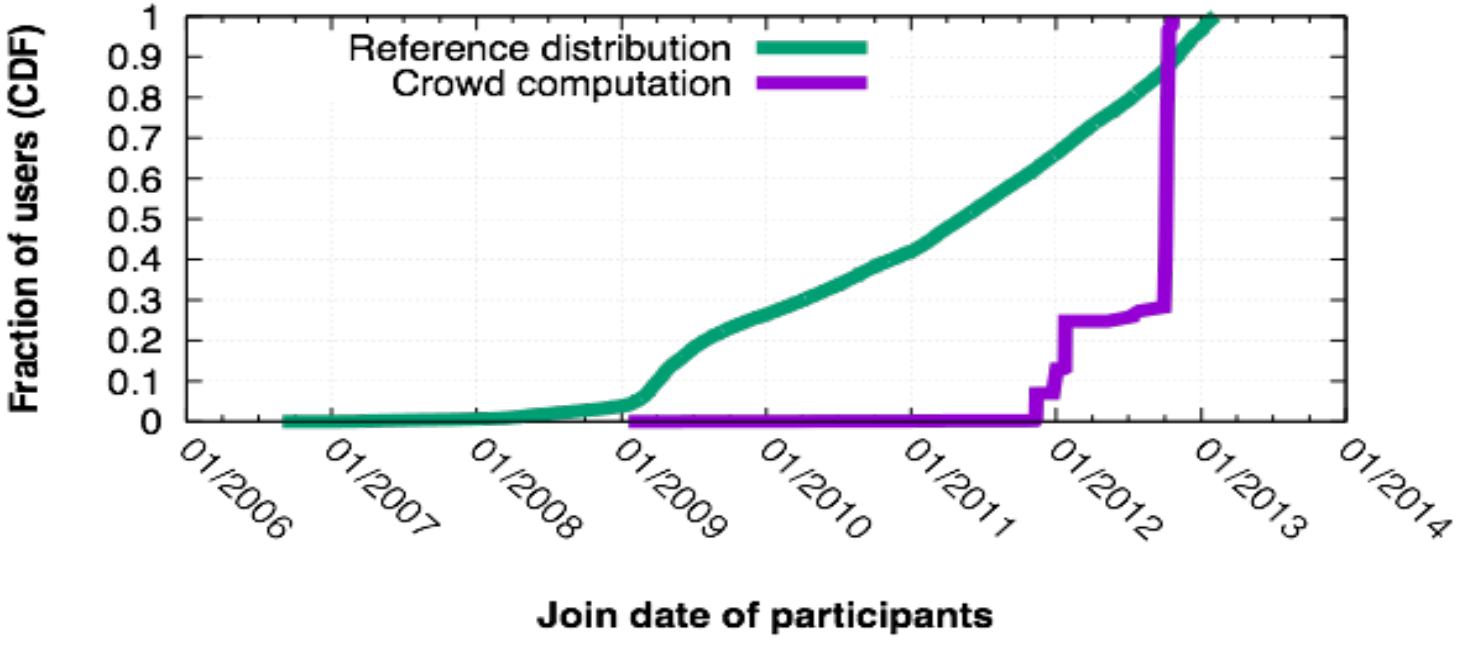


# Robustness against adaptive attackers

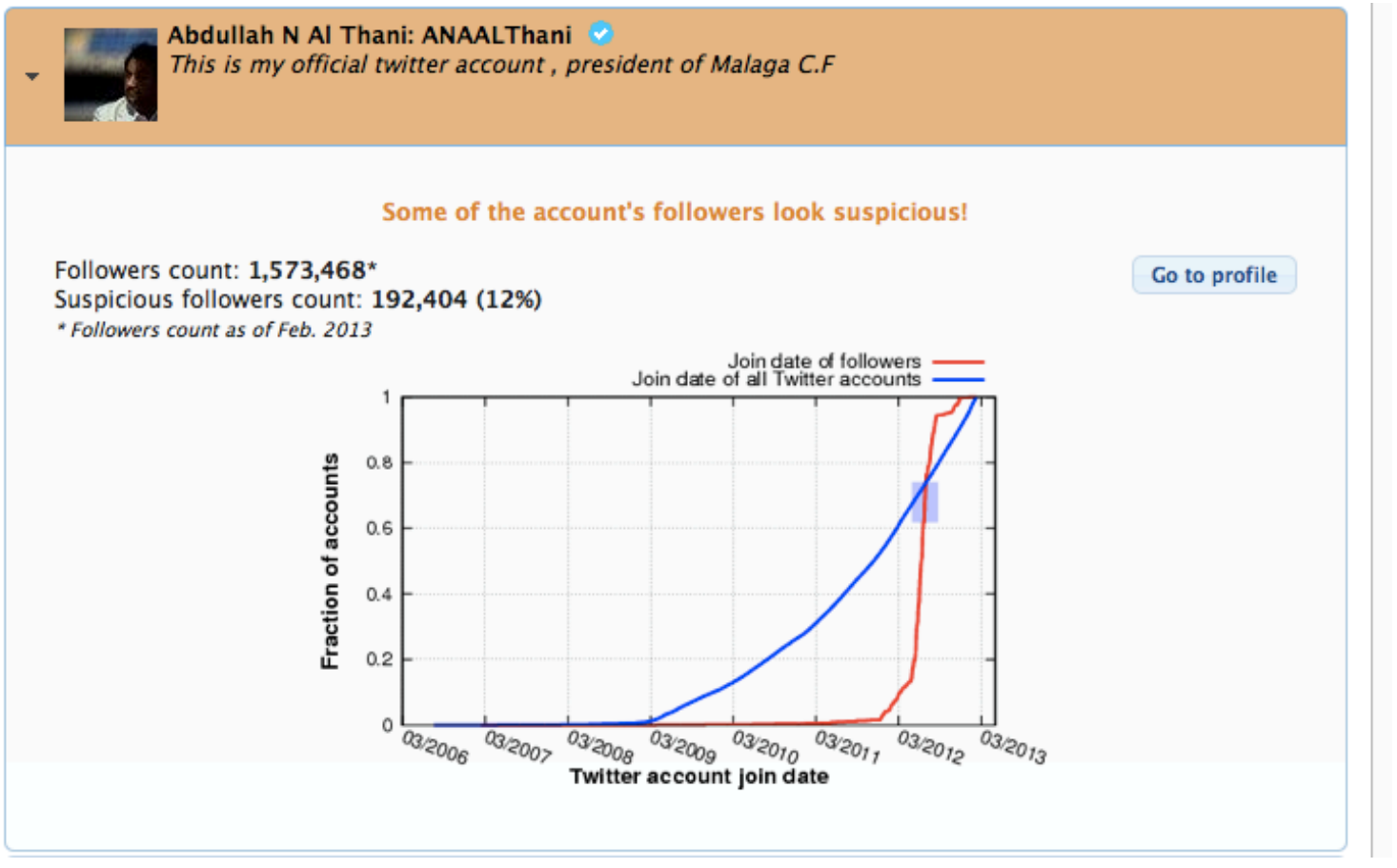
Stamper can fundamentally alter the arms race with attackers
What about attacks using compromised or colluding identities?

Compromised/colluding identities have to be selected in such a way that it would not affect the reference distribution

Any compromised identity that gets suspended leads to a near permanent damage to attacker's power



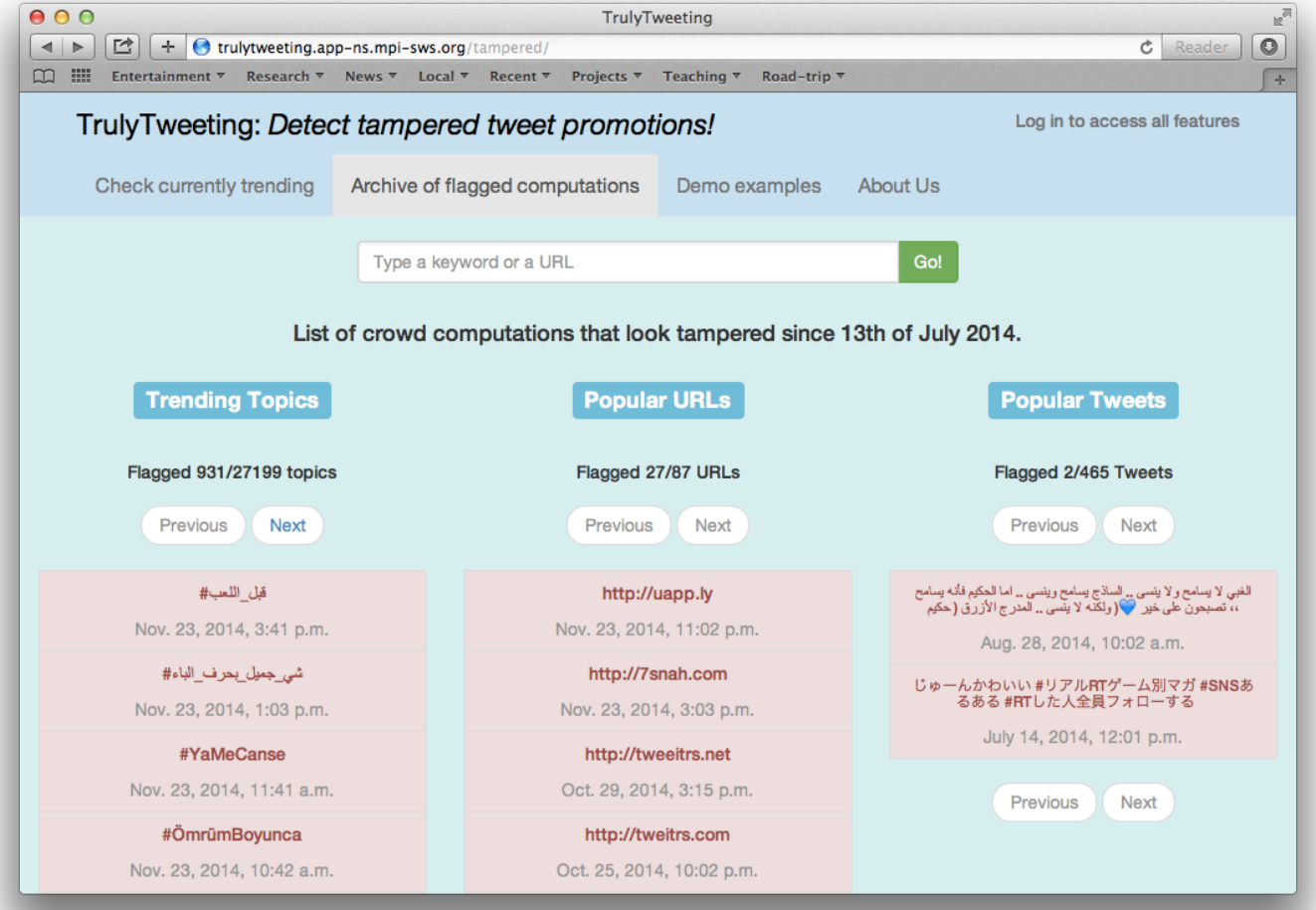
# TrulyFollowing: A prototype system




Detects popular users (politicians) with fake followers

- trulyfollowing.app-ns.mpi-sws.org

# TrulyTweeting: A prototype system




Detects popular hashtags, URLs, tweets with fake promoters

- trulytweeting.app-ns.mpi-sws.org

# DEMO

# Detection by Stamper: How it works

- Assume unbiased participation in a computation
  - The join date distributions for ids in any large-scale crowd computation must match that of a large random sample of ids on the site
- Any deviation indicates Sybil tampering
  - Greater the deviation, the more likely the tampering
  - Deviation can be calculated using KL-divergence
- Rank computations based on their divergence
  - Flag the most anomalous computations

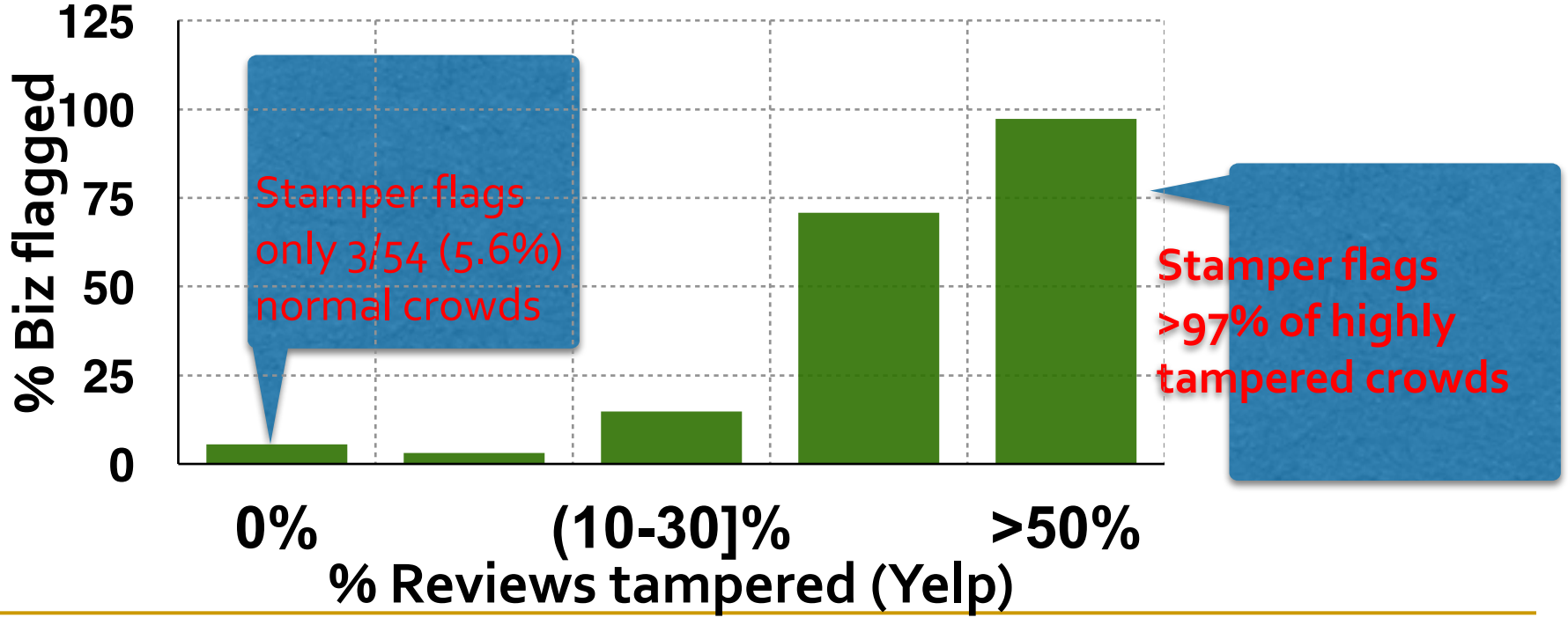
# Dealing with computations with biased participation

- When nodes come from a biased user population:
  - All computations suffer high deviations
    - Making the tamper detection process less effective
- Solution: Compute join dates' reference distribution from a similarly biased sample user population
  - I.e., select a user population with similar demographics
- Has the potential to improve accuracy further

# Detection accuracy: Yelp case study

- Case study: Find businesses with tampered reviews in Yelp
- Experimental set-up: 3,579 businesses with more than 100 reviews
  - "Ground-truth" obtained using Yelp's review filter

Stamper flags 362 businesses (83% of all with more then 30% tampering)



# Take-away lesson

- Ids are increasingly being profiled to detect Sybils
- Don't profile individual identities!
  - Accuracy would be low
  - Can't prevent tampering of computations
- Profile groups of ids participating in a computation
  - After all, the goal is to prevent tampering of computations

# Take-away questions

- What should a site do after detecting tampering?
- How do we know who tampered the computation?
  - Could a politician / business slander competing politicians / businesses by buying fake endorsements for them?
- Can we eliminate the effects of tampering?
  - Is it possible to discount tampered votes?

# Take-away questions

- In practice, users have weak identities across multiple sites
  - Such weak ids are increasingly being linked
- Can we transfer trust between weak identities of a user across domains?
  - Can Gmail help Facebook assess trust in Facebook ids created using Gmail ids?
- Can a collection of a user's weak user ids substitute for a strong user id?

# This talk

Explore temporal dimensions of

- ~~Trust in social computing systems~~
- Privacy in social computing systems