On the Strength of Weak Identities in Social Computing Systems: Or, How We Learnt to Reason about the Trustworthiness of Weak Identities

Krishna P. Gummadi

Max Planck Institute for Software Systems

Social computing systems

- Online systems that allow people to interact
- Examples:
  - 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
- Widely used & important

But, they have an achilles heel

- Users operate behind weak identities
  - Anyone can create an account
  - Fill in arbitrary profile information
  - No certification required from trusted authorities
    - E.g., passport, social security number, credit card

- Good: Preserves users' privacy / anonymity
  - In practice, many users provide offline identities
  - Some sites even require users to provide real names

- Bad: Vulnerable to Sybil (fake identity) 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

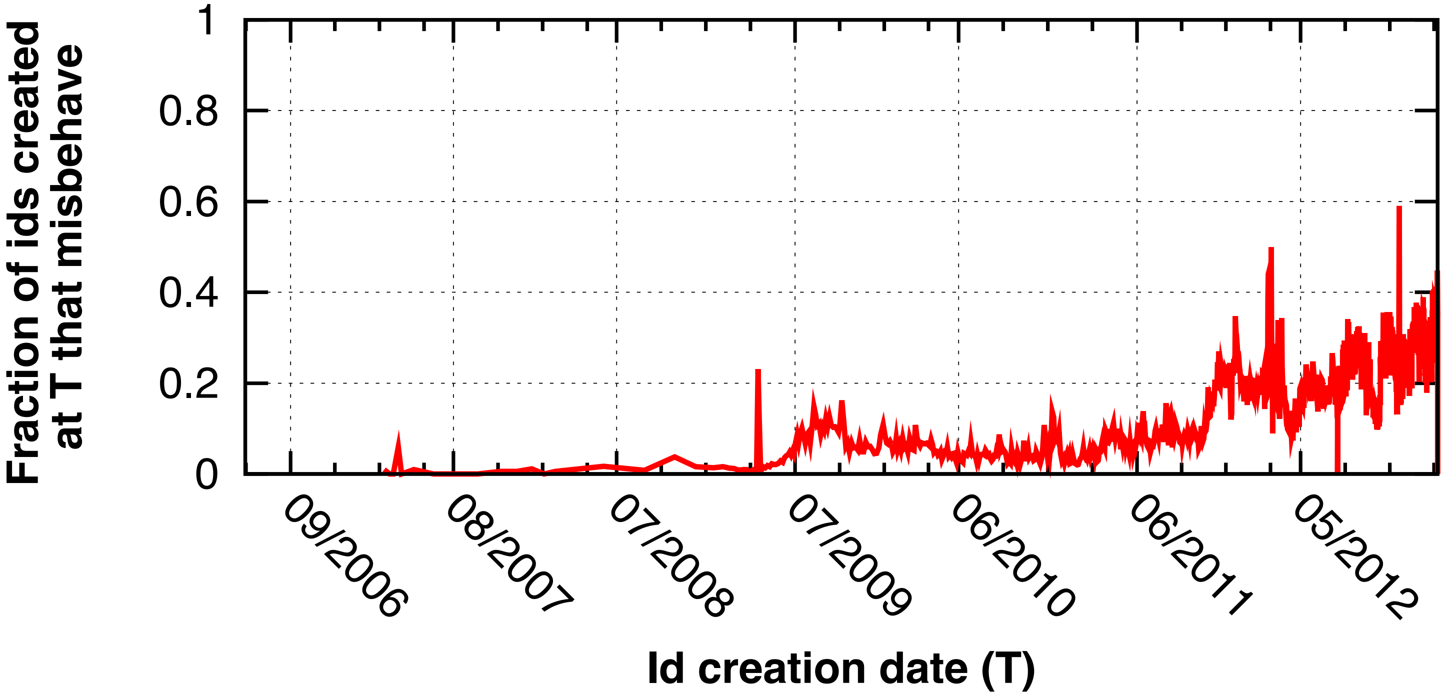
Instagram "Likes" Worth More Than Stolen Credit Cards

Posted by **samzenpus** on Monday August 19, 2013 @11:01AM
from the with-a-little-help-from-my-bots dept.

Barence writes

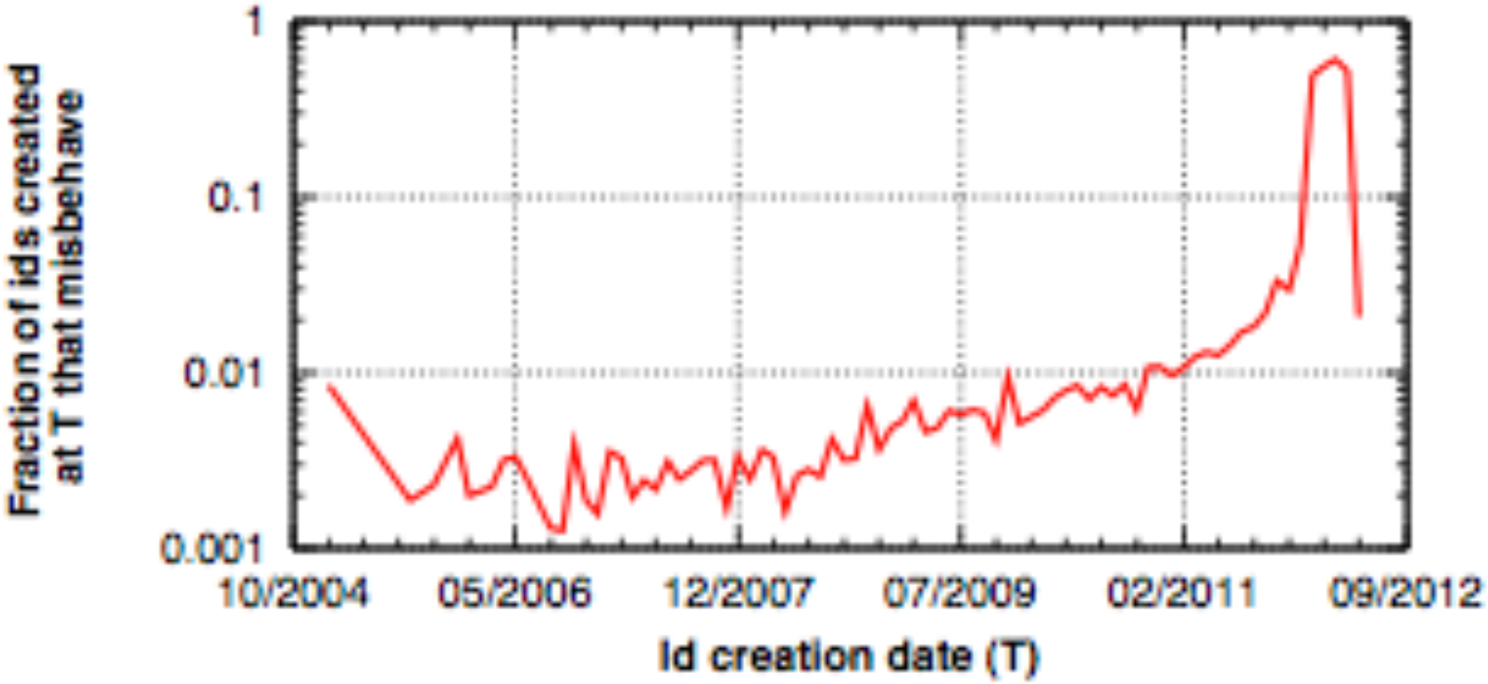
> "In the world of online fraud, a fake fan on Instagram can be worth five times more than a stolen credit card number. In a sign of the growing value of social network 'likes', the Zeus virus has been modified to create bogus Instagram 'likes' that can be used to generate buzz for a company or individual, according to cyber experts at RSA, the security division of EMC. These fake 'likes' are sold in batches of 1,000 on hacker forums, where cybercriminals also flog credit card numbers and other information stolen from PCs. According to RSA, 1,000 Instagram 'followers' can be bought for $15 and 1,000 Instagram 'likes' go for $30, whereas 1,000 credit card numbers cost as little as $6."

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!

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

This talk

- Explore limitations of existing approaches & ways to overcome them
- Part 1: Profiling user ids to detect Sybils
- Part 2: Leveraging social networks to detect Sybils

Part 1

Profiling user ids to identify Sybil ids

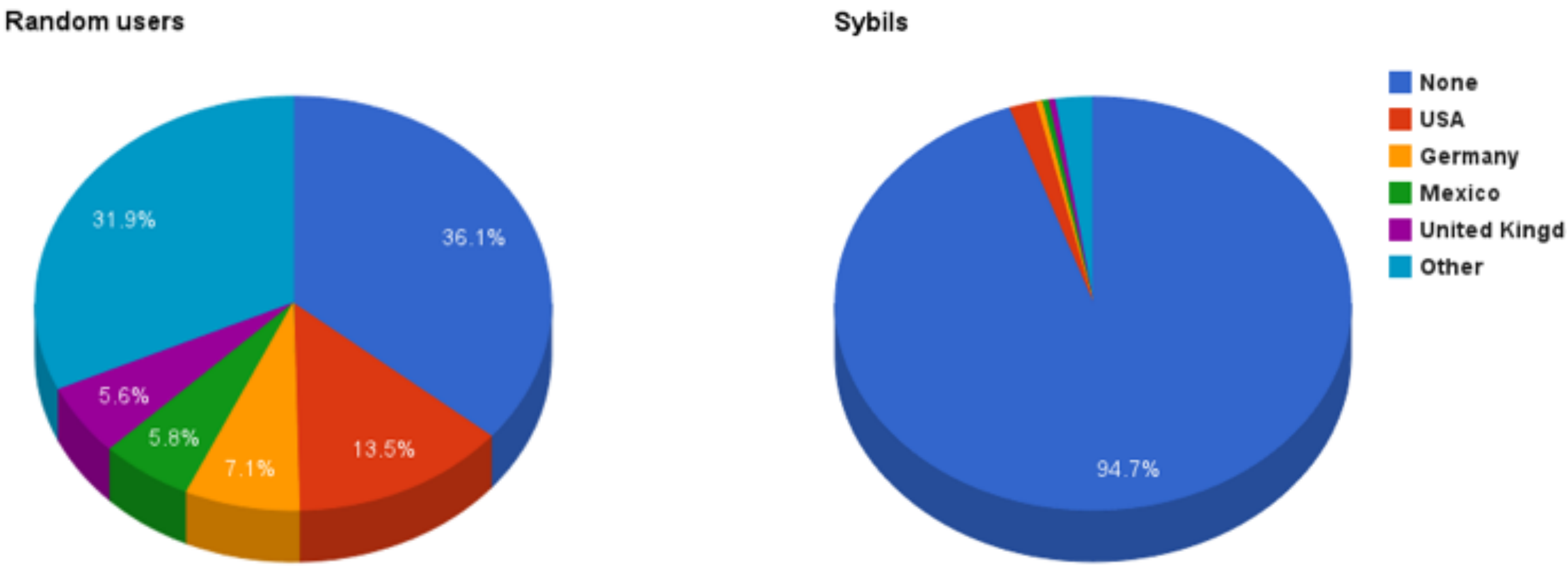
Lots of recent work

- 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

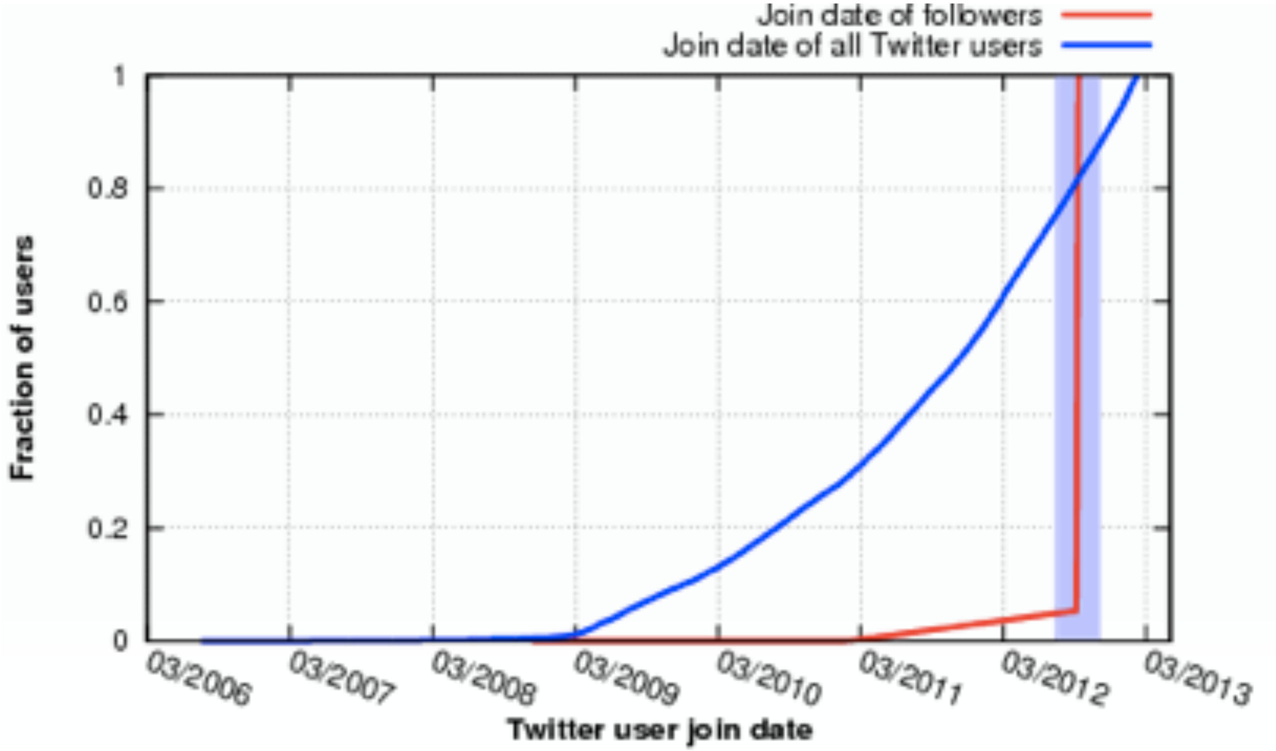
- For many profile attributes, the values assumed by Sybils & non-Sybils tend to be different
  - Location field is not set for >90% of Sybils, but <40% of non-Sybils
  - Lots of Sybils have low follower-to-following ratio
  - A much smaller fraction of Sybils have more than 100,000 followers

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

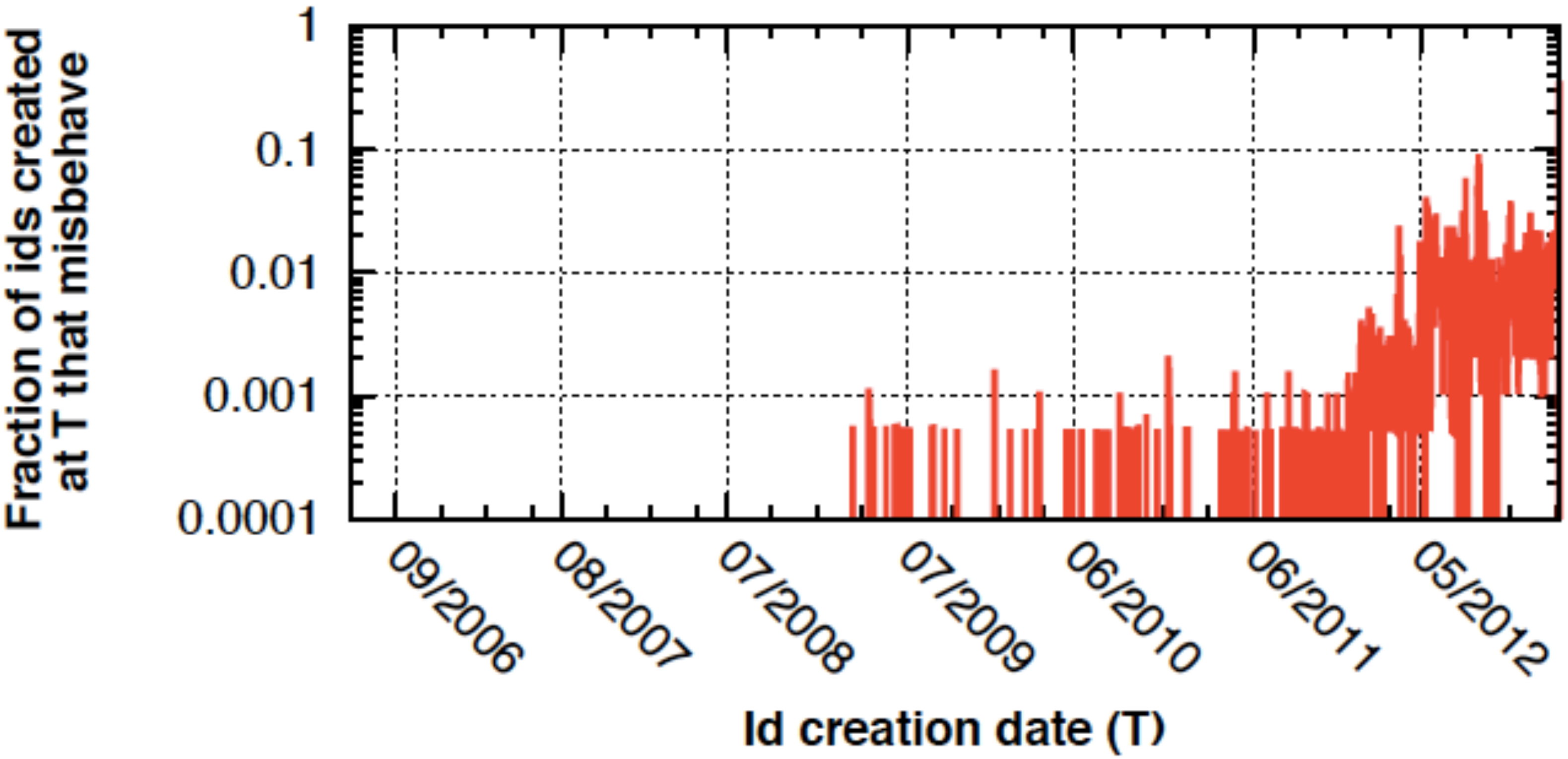


- All our bought fake followers were newly created!

Two key observations

- Attackers cannot tamper their join dates (id creation timestamps)
- Older ids are more trustworthy 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 more trustworthy than newer ids

Robust tamper detection in crowd computations

- Insight: Can detect tampered computations even when we cannot detect fake ids
- Idea: Detect tampering by analyzing join date distributions of participants
  - Entropy of tampered computations tends to be lower
- Approach is robust against adaptive attackers
  - Attackers have to create ids from the system's inception
  - Attack power decreases with every suspended id

Our Stamper project

- Profile crowd computations, not individual ids
  - Profile the set of ids involved in a common activity
    - E.g., rating a restaurant, following a user, promoting a tweet

- Assuming unbiased participation, the join date distributions for ids in any large-scale crowd computation must match those for honest ids

- Any deviation indicates Sybil tampering
  - Greater the deviation, the more likely the tampering
  - Deviation can be calculated using KL-divergence

DEMO

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 significant potential to improve accuracy further

Take-away lesson

- Identities 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

This talk

- Explore limitations of existing approaches & ways to overcome them
- Part 1: Profiling user ids to detect Sybils
- Part 2: Leveraging social networks to detect Sybils

Part 2

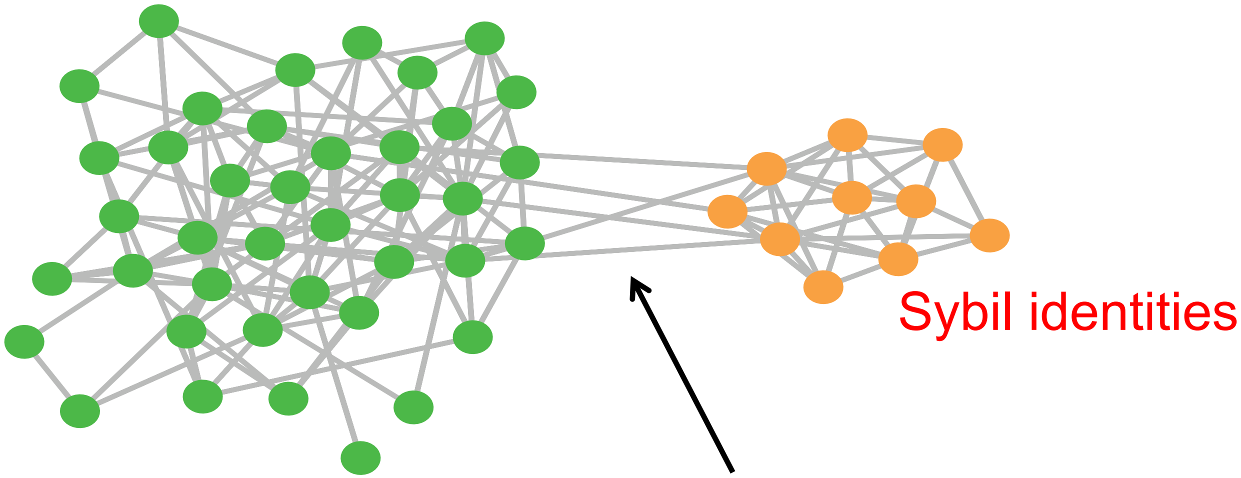
Social network-based Sybil id detection

High-level idea

Assumption: Links take some effort to form and maintain

E.g.: Good users only accept links from users they recognize

Assumption holds in some though not all social networks



Attacker is limited by his ability to form social links to real users

Lots of recent work

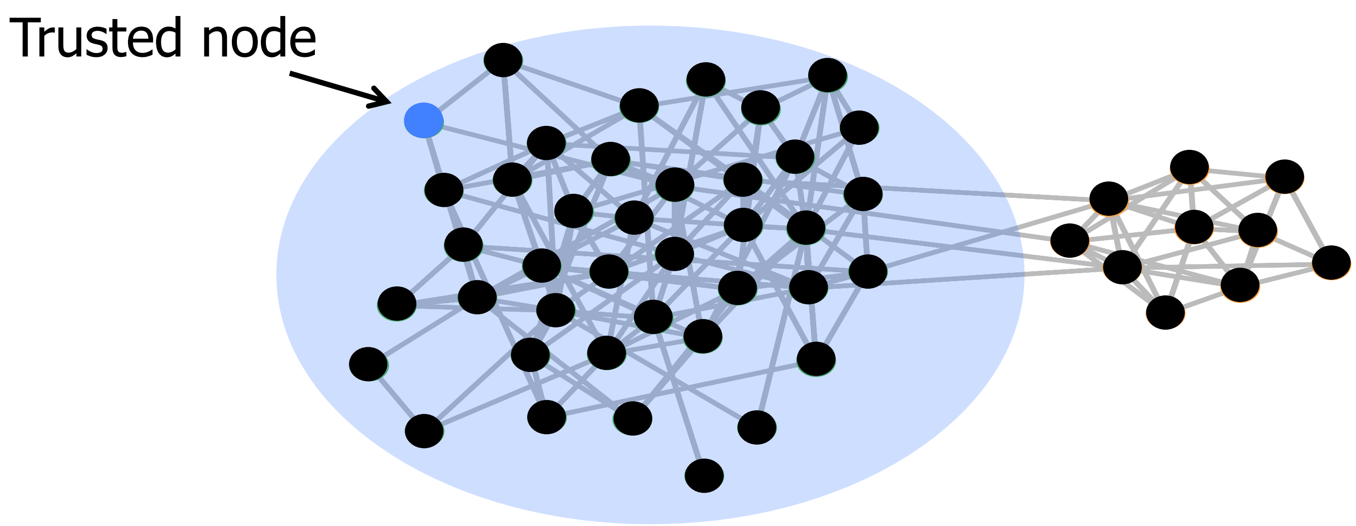
- Sybil detection: Identify Sybil nodes & block
  - *SybilGuard [SIGCOMM '06], SybilLimit [Oakland S&P '08], SybilInfer [NDSS '08], MOBID [INFOCOM '10], GateKeeper [INFOCOM '11], SybilRank [NSDI '12]*
- Model: Given a social network & at least one non-Sybil node, they identify Sybil identities
  - By analyzing only the network's graph structure

How Sybil detection works

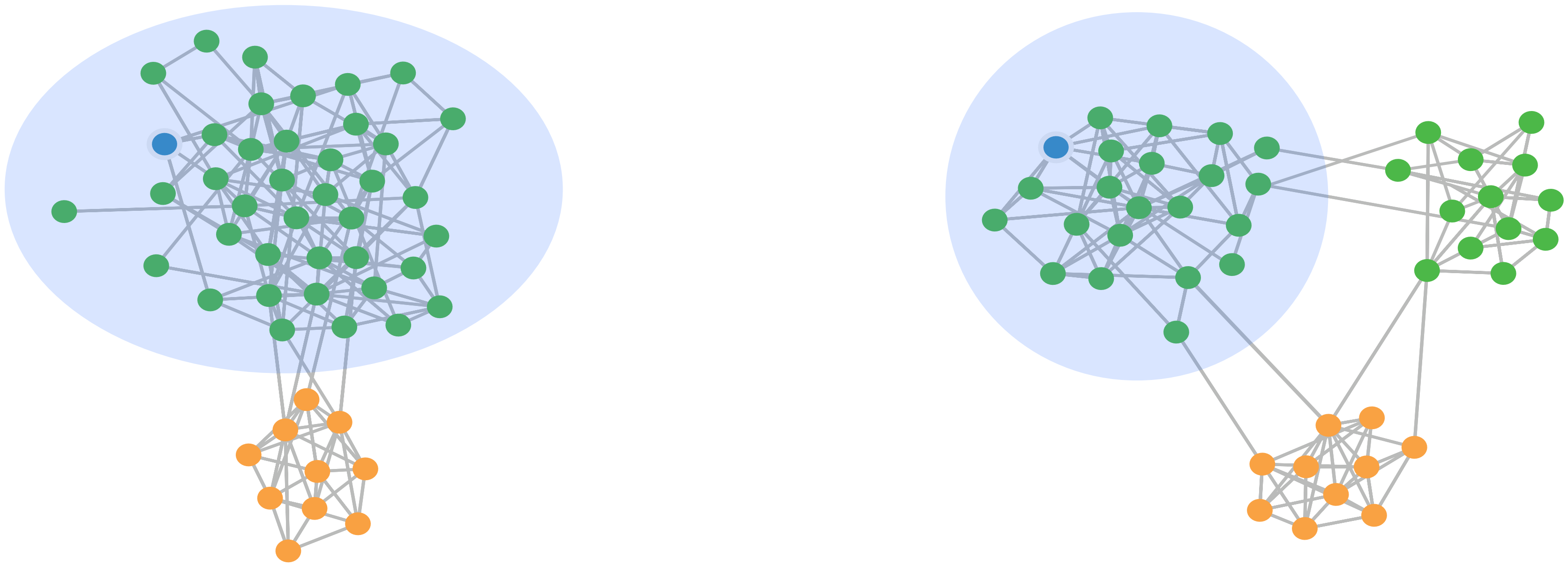
- All algorithms perform random walks from *a priori* trusted nodes
  - The exact nature of random walks differ
- Nodes are ranked based on their *closeness* to the trusted nodes *[SIGCOMM '10, NSDI '12, Oakland S&P '13]*
  - Nodes that have a higher chance of being visited are ranked closer
  - Very similar to TrustRank on Web graph [VLDB '04]
- Nodes beyond a threshold rank are declared Sybils

Key challenge in practice

- Picking threshold rank separating Sybils & non-Sybils
- A good demarcating threshold exists, only when
  1. The non-Sybil network is fast mixing (tightly-knit)
  2. The Sybil network has limited connectivity to non-Sybils



Do non-Sybils form a single, tightly-knit community?



- Large-scale social nets have small fringe communities *[Leskovec 2008], [Dell'Amico 2009]*

- Sybil clouds and small communities would be indistinguishable using the graph structure alone

Sybil detection in practice

- Cannot pick a good threshold to blacklist Sybil ids
  - To date, no scheme has been applied in practice
- But, we can conservatively white-list non-Sybil ids
  - Nodes that are ranked close to the trusted nodes

Our Trusty project

- Goal: Finding trustworthy content in Twitter micro-blogging site

- Key ideas:
  - Twitter has over 50K *a priori* verified ids
  - Use them to propagate trust in the Twitter network graph
  - White-list as many Twitter network ids as possible
  - Tweets from white-listed ids would be more trustworthy than tweets from random ids

Challenge: Link farming in Twitter

- Many popular and verified identities reciprocate follow-links from arbitrary nodes [WWW '12]

- Follow-links in Twitter do not necessarily imply trust

- Propagating trust on Twitter follow network spreads trust to spammers as well

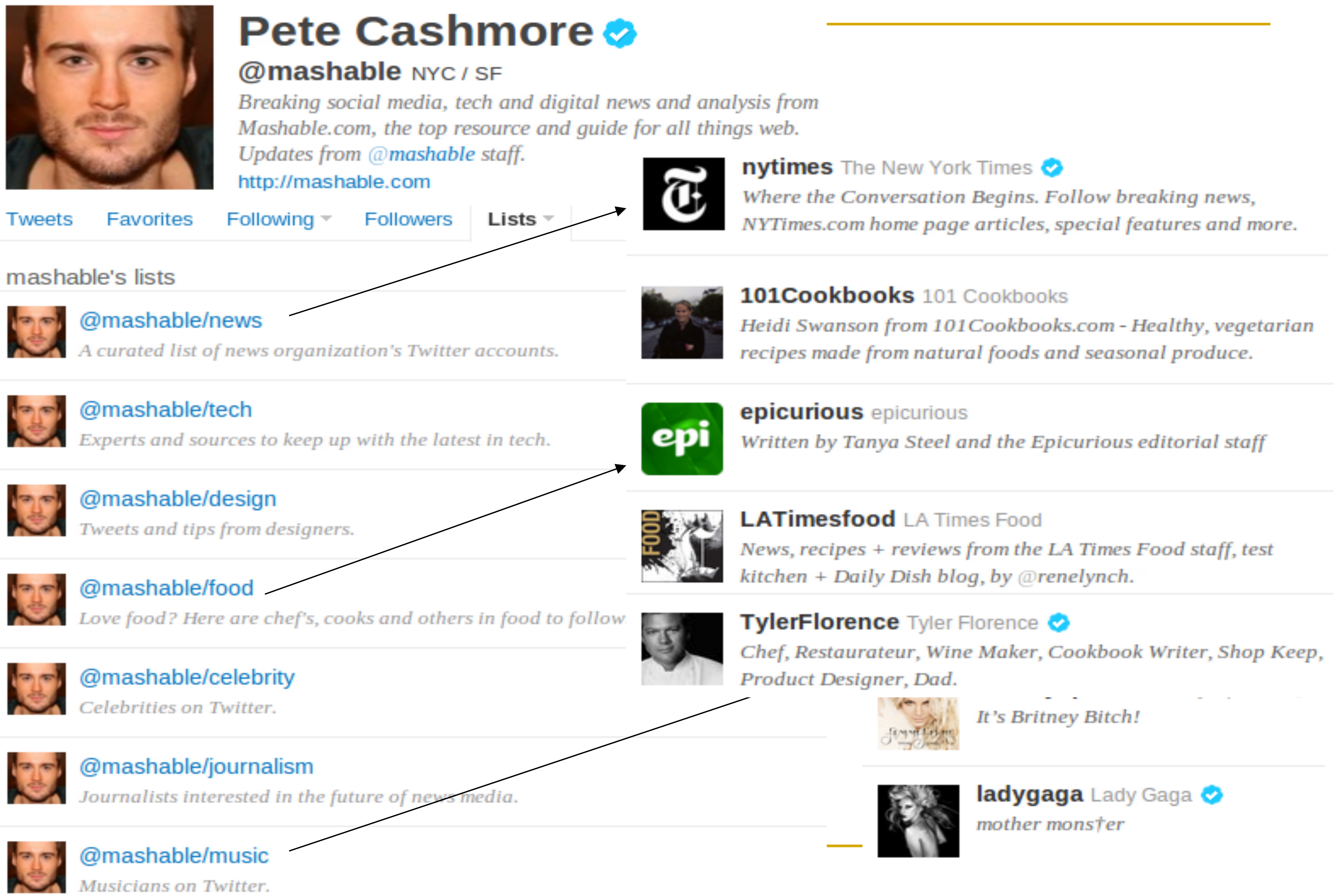
How to infer trust between ids in Twitter?

Inferring trust between Twitter ids

- Twitter Lists: A feature to organize tweets received from the people whom a user is following

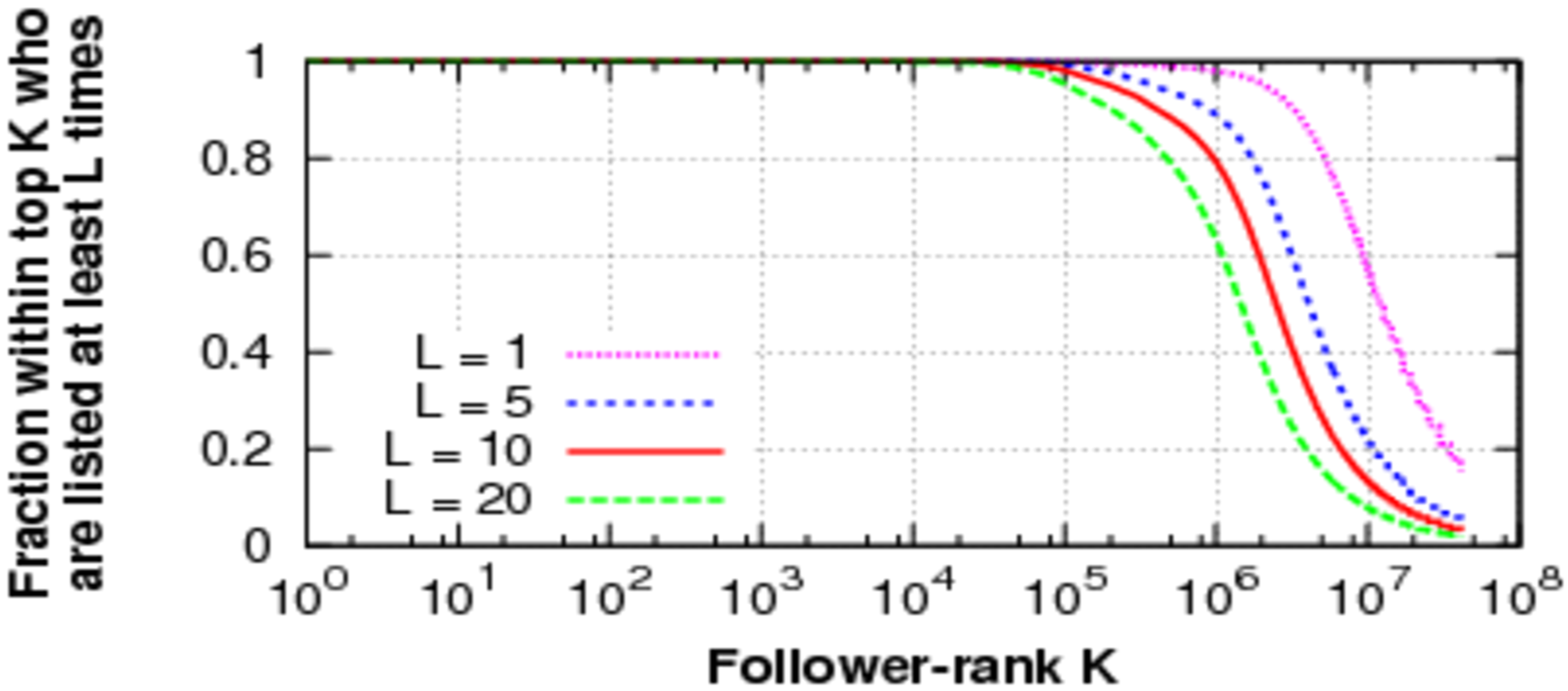
- Create a List, add name & description, add Twitter users to the list
  - List meta-data offers cues for who-is-who
  - Tweets from listed users appear in a separate List stream

- Insight: Good users don't list spammers as experts
  - Even when they follow them


Pete Cashmore
@mashable NYC / SF
Breaking social media, tech and digital news and analysis from Mashable.com, the top resource and guide for all things web. Updates from @mashable staff.
http://mashable.com
Tweets
Favorites
Following
Followers
Lists
mashable's lists
@mashable/news
A curated list of news organization's Twitter accounts.
@mashable/tech
Experts and sources to keep up with the latest in tech.
@mashable/design
Tweets and tips from designers.
@mashable/food
Love food? Here are chef's, cooks and others in food to follow
@mashable/celebrity
Celebrities on Twitter.
@mashable/journalism
Journalists interested in the future of news media.
@mashable/music
Musicians on Twitter.
nytimes The New York Times
Where the Conversation Begins. Follow breaking news, NYTimes.com home page articles, special features and more.
101Cookbooks 101 Cookbooks
Heidi Swanson from 101Cookbooks.com - Healthy, vegetarian recipes made from natural foods and seasonal produce.
epicurious epicurious
Written by Tanya Steel and the Epicurious editorial staff
LATimesfood LA Times Food
News, recipes + reviews from the LA Times Food staff, test kitchen + Daily Dish blog, by @renelynch.
TylerFlorence Tyler Florence
Chef, Restaurateur, Wine Maker, Cookbook Writer, Shop Keep, Product Designer, Dad.
It's Britney Bitch!
ladygaga Lady Gaga
mother monster

What fraction of users are Listed?

[WOSN '12, SIGIR '12]



Overall, 2.5% of all Twitter users are *Listed*

But, an overwhelmingly large fraction of popular nodes are *Listed*

White-listing nodes in Twitter

- Can run TrustRank on List-network
  - Starting with verified Twitter users as seed set
- Ran TrustRank over the network of List-links
- Conservatively, white-listed all nodes that lie within top-third of trusted nodes

Is content from white-listed users trustworthy? *[CIKM '13]*

- Analyzed tweets from white-listed users for spam
  - Compared with a similarly-sized set of random tweets from all Twitter users

- Tweets from white-listed users have an order of magnitude fewer spam tweets than random sample

- Better still, they are rich in information content as they are from authoritative topical experts

DEMO

Take-away lesson

- Social networks can be used for propagating trust
- In practice, they are more effective at whitelisting non-Sybil nodes
  - Not for blacklisting Sybil nodes!
- Lots of practical applications

Summarizing the take-away lessons

- Don’t profile individual identities
  - Profile groups of ids participating in a computation
- Don’t use social links (trust) to blacklist Sybils
  - Use social trust (links) to whitelist non-Sybils