DIG Student Seminar

Spamming on the Social Web

Albert Au Yeung

12 November 2008

Agenda

- Problem of spamming on social Web sites
- Types of anti-spamming strategies
- Evaluation of anti-spamming techniques
- Spamming in collaborative tagging

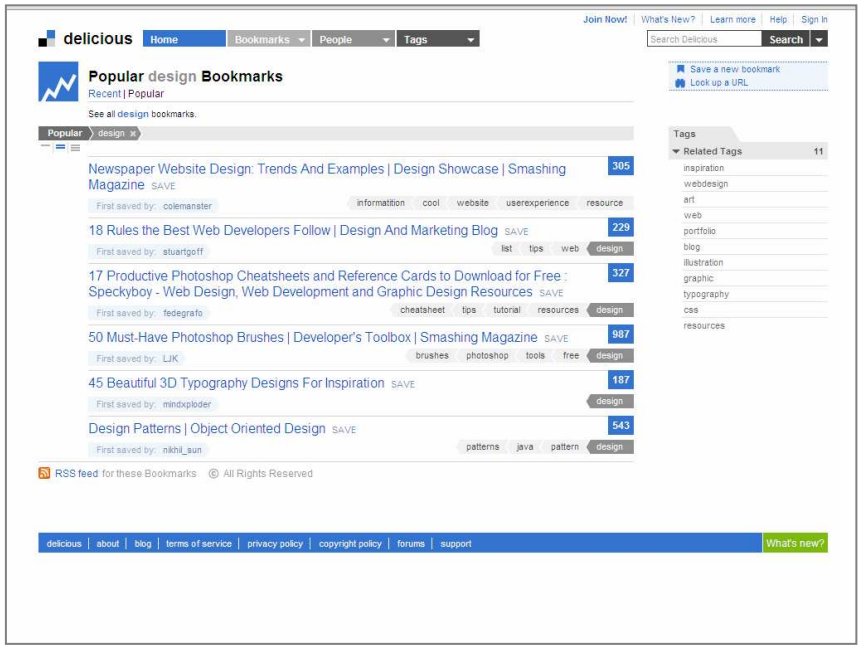
Spamming

- Social Web sites: content is primarily supplied by general users
- New means of sharing/discovering useful/interesting resources on the Web
- Also interested by spammers: new means of attracting user attentions
- Spam hinders resource sharing and discovery (e.g. 19 of 20 top most active users in Delicious are spammers (Wetzker et al. 2008))

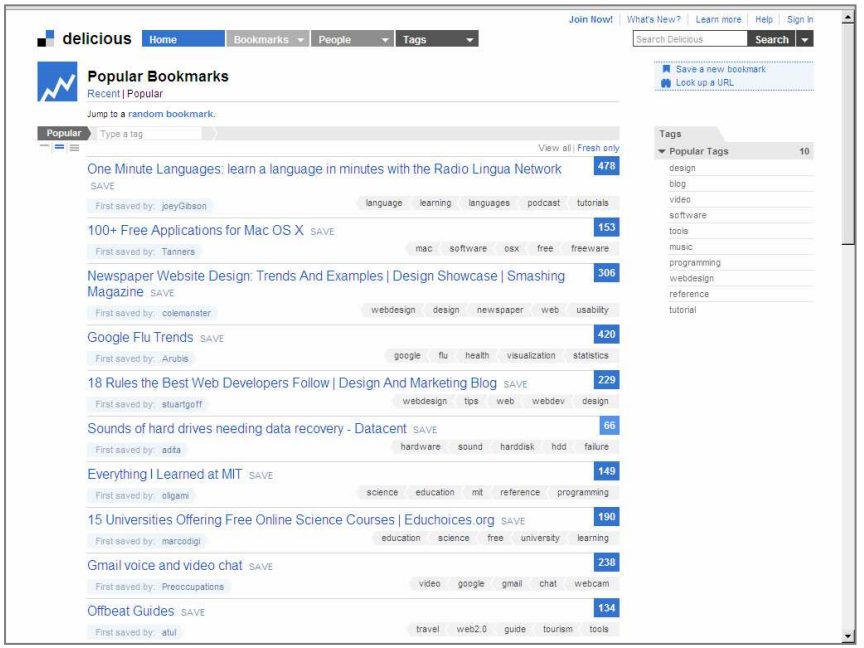
Characteristics of Social Web Sites

- Centralisation:
  one controlling entity, runs its own servers, has access to all the data
- Well-defined interactions:
  user behaviours tend to be more predictable
- Identity:
  every user's action can be traced
- Multiple interfaces:
  more chances to get spammed, more different anti-spamming techniques are required

Example: Delicious

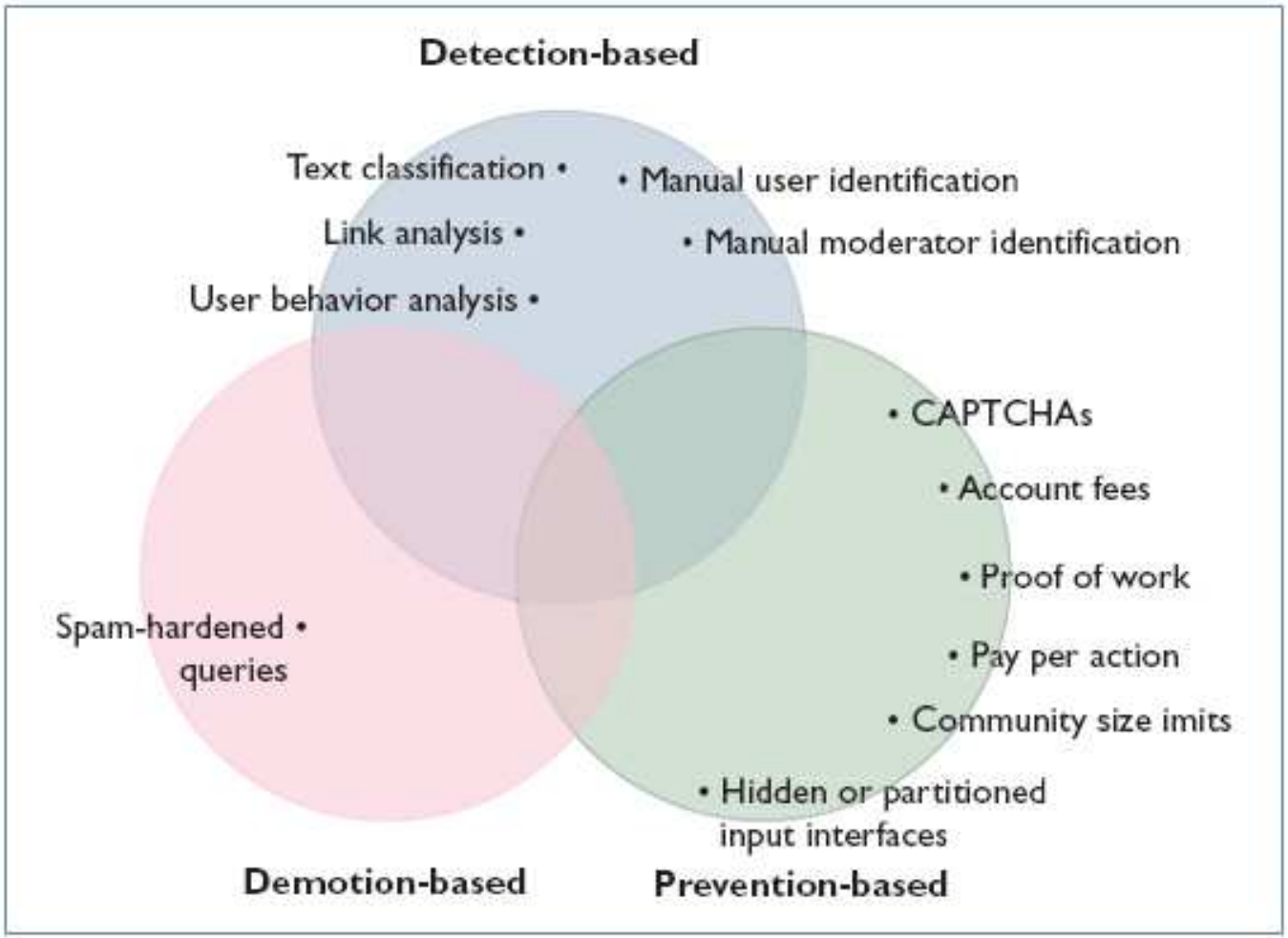


Browsing bookmarks with a certain tag



Browsing popular bookmarks

Anti-spamming Strategies

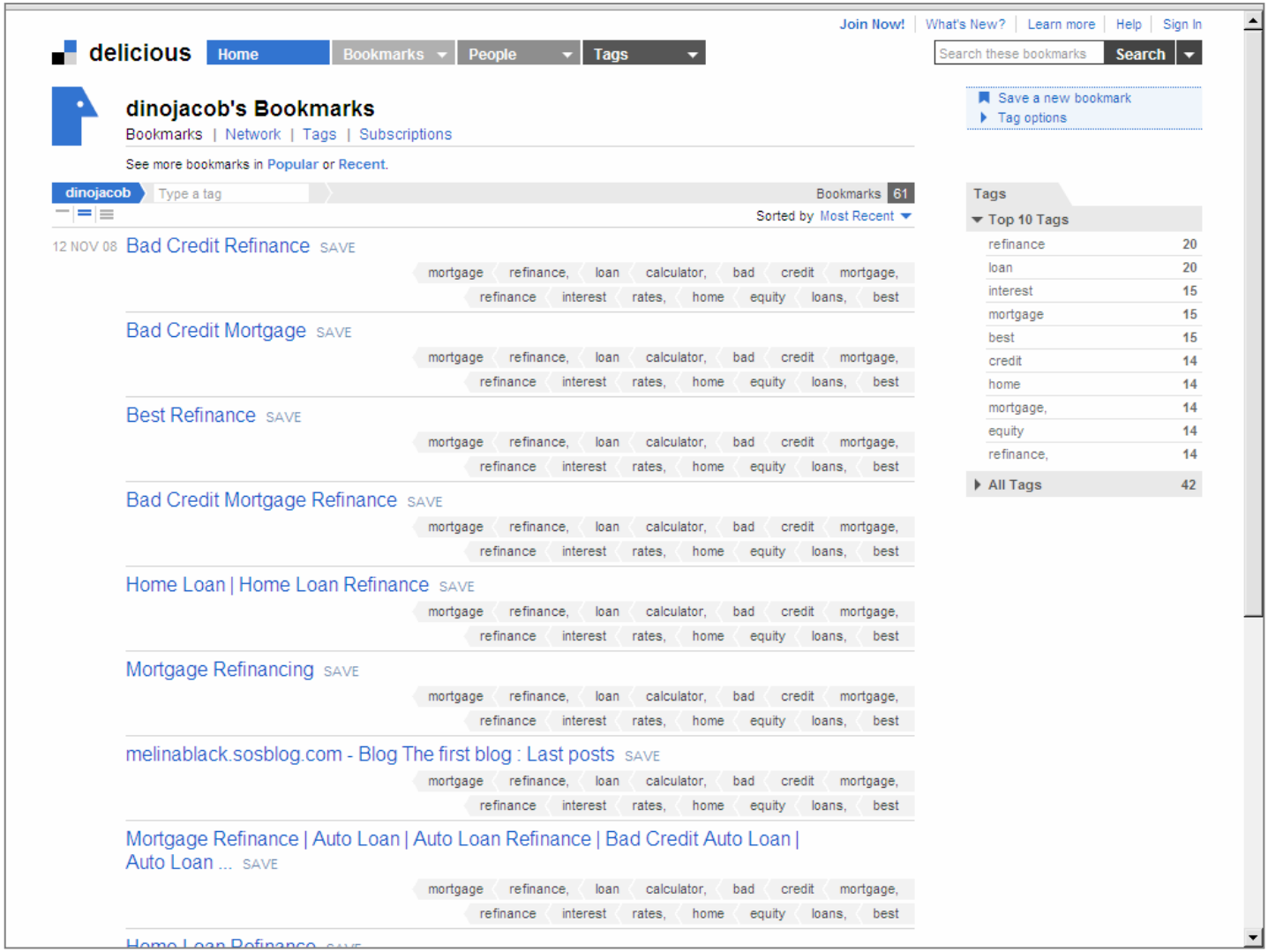


Anti-spamming Strategies

Detection-based Anti-spamming

- Simplest form: manual identification
- Pattern-based automatic identification
- Usually involves supervised learning (need training data or seed data)
- Example in tagging: common patterns include
  (1) bookmarks within the same domain
  (2) use title words as tags
  (3) use same tags across many bookmarks

Anti-spamming Strategies



Anti-spamming Strategies

Demotion-based Anti-spamming

- Reduce the prominence of spams (remove it from the top of the lists)
- Ideally providing ranking which promotes good content and demotes spams at the same time
- With a good algorithm, this should be the most desirable form of anti-spamming
- Need different algorithms for different interfaces (e.g. most recent, most popular, single user/tag)

Anti-spamming Strategies

Prevention-based Anti-spamming

- Making spamming activities more difficult to carry out
- Example: using CAPTCHA

- Make contributing content cost computational time or even real money (e.g. the PennyBlack project at Microsoft)
- Problem: causing inconvenience to real users at the same time

Evaluation of Anti-spamming

Spam Models

- Give a model of the spamming activities on a site
- Describe what a spam would look like
- (1) synthetic model
  - representation of a social system
  - definition of spamming (e.g. wrong tags)
  - may not capture activities of real spammers
- (2) trace-driven model
  - require labelling of real data
  - produce more realistic results
  - but may result in bias, also time-consuming

Evaluation of Anti-spamming

Spam Metrics

- Quantify the impact of spam on an interface
- Measure effectiveness of a anti-spamming technique
- Examples:

  (1) Precision & Recall

  (2) SpamFactor
  (Koutrika et al. 2007)

$$SpamFactor(n,t) = \frac{\sum_{i=1}^{n} w(o_i,t) * \frac{1}{i}}{\sum_{i=1}^{n} \frac{1}{i}},$$

where

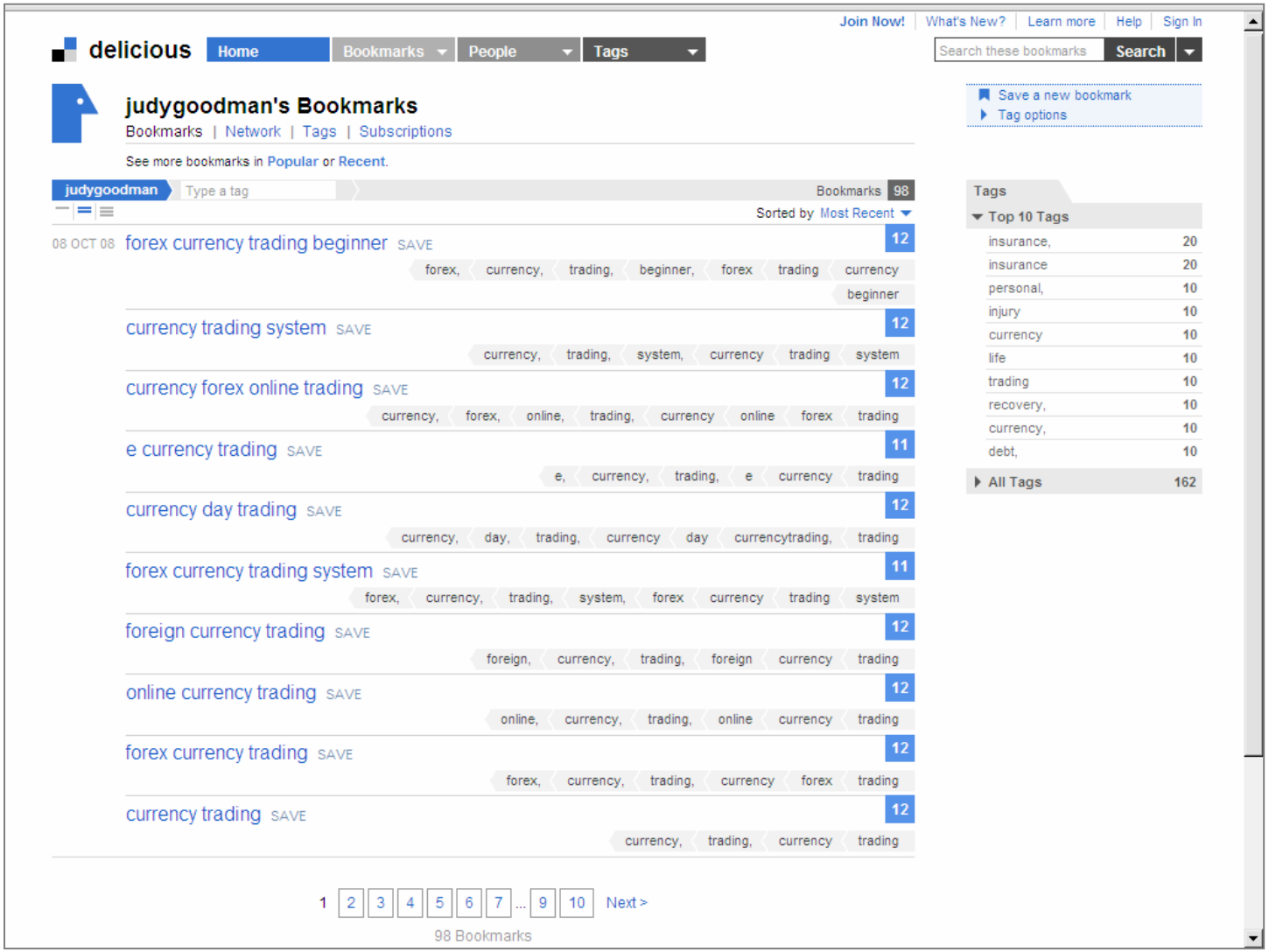
$$w(o_i,t) = \begin{cases} 1 \text{ if } t \text{ is a bad tag for } o_i \\ 0 \text{ if } t \text{ is a good tag for } o_i \end{cases}.$$

Spamming in Collaborative Tagging

Types of spammers

- *Flooders*
  - Tag a large number of existing documents
  - Aim at gaining reputation/high ranks
- *Promoters*
  - Contribute a lot of new documents (spam Websites)
  - Aim at promoting their own documents
- *Trojans*
  - Disguise themselves by contributing popular bookmarks
  - Direct attention of users to malicious Web sites

Spamming in Collaborative Tagging

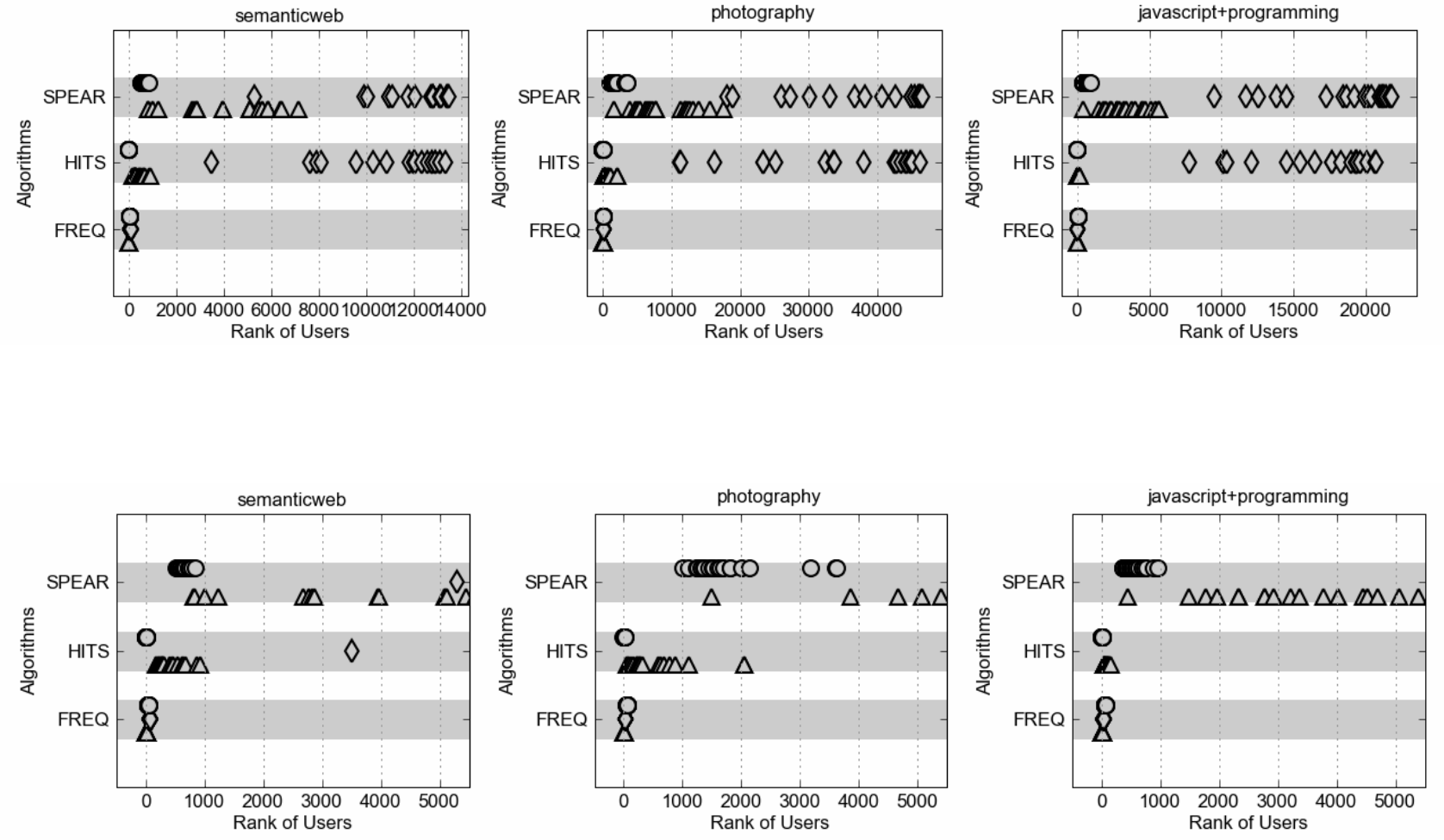


Spamming in Collaborative Tagging

Spamming-resistant Ranking

- HITS-like algorithm
  - Good and expert users have high quality bookmarks
  - High quality bookmarks are tagged by good users
- Discoverer/Follower Scoring Scheme
  - Discoverers are credited more
  - Followers receive lower score
- Experiments show good results against simulated spammers

Spamming in Collaborative Tagging



Conclusions

- Spamming hinders resource sharing and retrieval on social Web sites
- Spamming on social Web sites comes in different forms, has more targets
- Future anti-spamming techniques need to focus on:
  (1) mixing different strategies
  (2) community-specific method