Effort-based Detection of Comment Spammers

Acar Tamersoy, Georgia Tech
Hua Ouyang, Yahoo Labs
houyang@yahoo-inc.com
Polo Chau Georgia Tech

Summary

We tackle the crucial problem of **comment spam** and propose **EDOCS**, a graph-based approach that quantifies how much ***effort*** a user exerted over his or her comments, to detect if the user is a comment spammer or not. EDOCS is effective in detecting comment spammers accurately with **95% true positive rate** at **3% false positive rate** as well as **preemptively**.

Who are Comment Spammers?

Comment spammers use comment treads to post irrelevant content (**spam**). A recent study showed that over **75% of the one million** blog comments collected were indeed spam, some with links to **malware** sites.

Spammers are Smart!

Spam comments are often short and carefully crafted. **Even human experts have a hard time differentiating** some spam comments from legitimate ones.

A Real-World Example

Original post:
Recently I signed up with walters alerts “Google em” they sounded pretty good in there emails about there picks, so I decided to give them a shot and bought there last pick VISN at $2.40 boy, was I amazed I ended up selling for 300% profit.

Several replies labeled as “clean” by human editors:
Re: Great, i got some shares yesterday. Good luck.
Re: Re: FACTS!!! I love it ! I agree.
Re: Re: Re: good posts need to be at the top ...

Why Quantify Effort to Detect Spammers?

Intuitively, **spammers would only exert limited time** and **money** when preparing and disseminating comments.

Our *Effort-based Detection of Comment Spammers* (**EDOCS**) algorithm captures this intuition, by analyzing a bipartite graph of users and effort-related feature values to **quantify how much effort a user exerted** over his or her comments.

Effort scores of comment spammers should be **lower** than those of the legitimate users.

A Graph-based Algorithm: EDOCS

EDOCS operates on a **bipartite graph of users and effort-related feature values** and performs **iterative message propagation** on this graph.

A **user is connected to all the feature values** that apply to her (e.g., connecting a user with her IP address).

We currently consider two features:

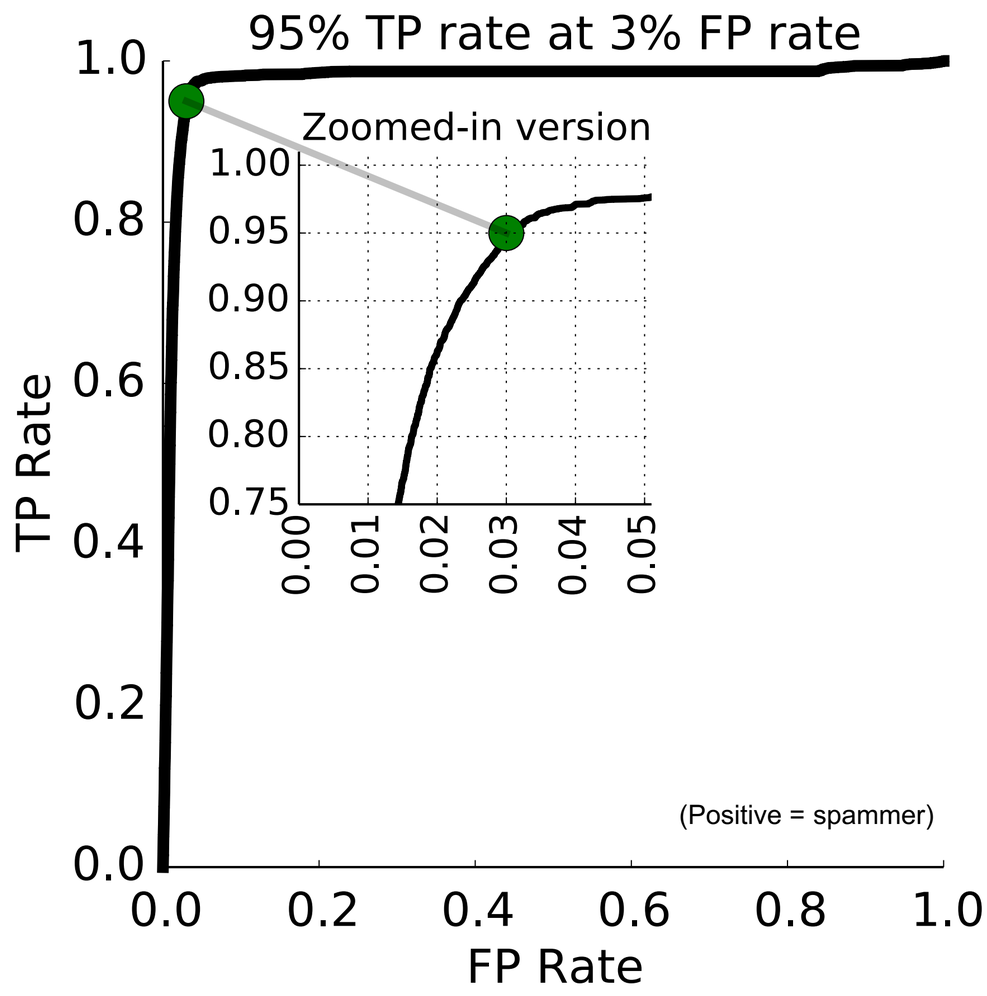
1. Effort to write comment text
2. Effort to obtain IP addresses

Yahoo Finance Dataset

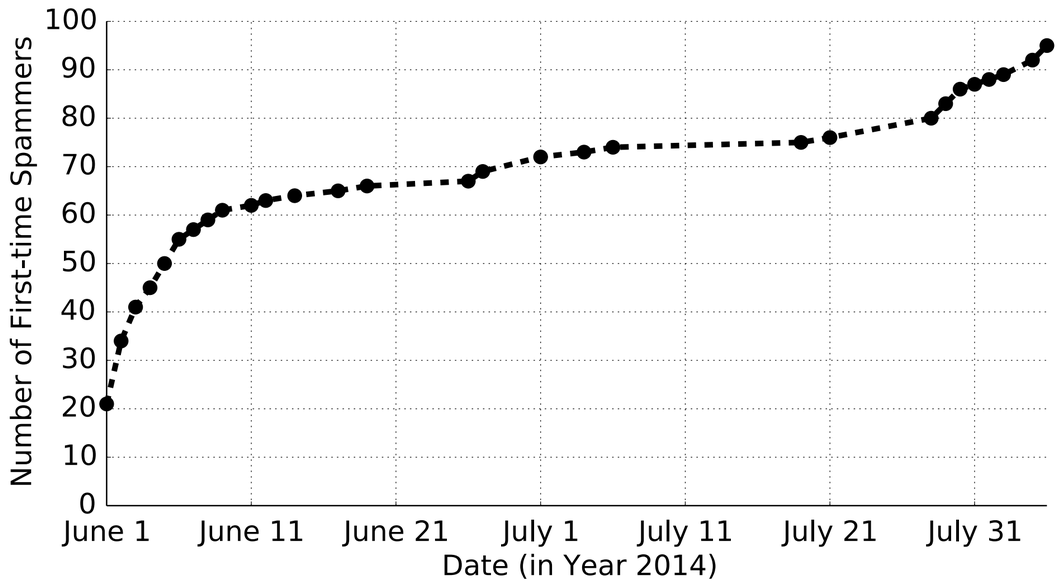
#users	197,464 (20.03% spammers)
#comments	1,201,277
Mean/Median #comments per user	6.08/1
Dataset duration	May 1-31, 2014
Duration of follow-up period	June 1-August 5, 2014

A user is assumed to be a **spammer** if she posted at least one comment labelled as spam by human experts.

Experiments: Labelling Spammers (over 197k users)



Experiments: Follow-up on false alarms



Conversion trend of users from “clean” to spammer based on the date of their first spam comment messages during the follow-up period. EDOCS **preemptively detected these 95 users** (top right corner) as spammers using data from May 2014.