Think before RT: An Experimental Study of Abusing Twitter Trends

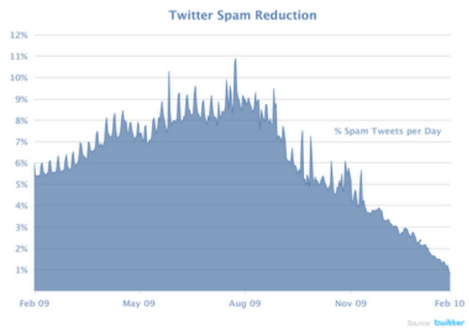
Despoina Antonakaki[1,2]
FORTH-ICS[1], University of Crete[2] <despoina@ics.forth.gr>

Introduction and objectives

twitter is a fast growing microblogging social network with more than 650M users. Messages cannot exceed 140 characters and may contain #HASHTAGS, @mentions and URLs. Users have followers and followings.

Popular Trends (PTs) are Hashtags that are popular in a region (i.e. Greece). An example of recent PTs are: #WorldCup14, #Gaza

Spam in Twitter has higher click-through rate than e-mail spam. Twitter has a URL shortening service that checks preemptively for spam. This reduces spam the rate to 1%.



Get More Followers (GMF) campaigns are a new class of Twitter Spam that escapes detection. It promotes tweets that lure users to give their credentials in exchange for more followers. This is multi million dollar scheme! They exploit (a) Google Results for obfuscating spam links and (b) unrelated Popular Trends.

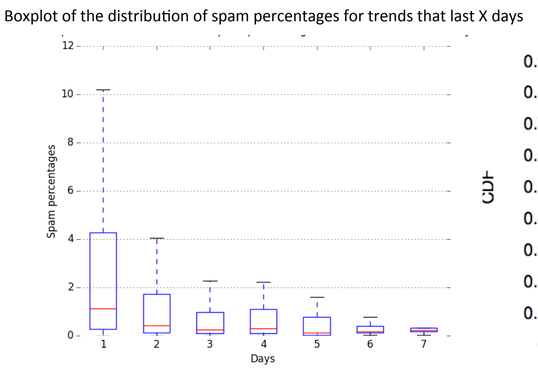
*www.google.com.tr/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=4&cad=rja&sqi=2&ved=0CEAQFjAD&url=http%3A%2F%2F**www.followmania.us**%2F&ei=FveHUoDBHITdswa2q4HYCw&usg=AFQjCNG1fWSqWsWxdI2QCMSeu3WGXHEtaw&sig2=fEuJc66pEvdq9vLp2XyGdg&bvm=bv.56643336,d.Yms …*

The Objectives of this study are: **(1)** Study the nature of exploited PTs and other Twitter features, **(2)** build a classification mechanism that discriminates GMF spam and **(3)** visualize GMF campaigns

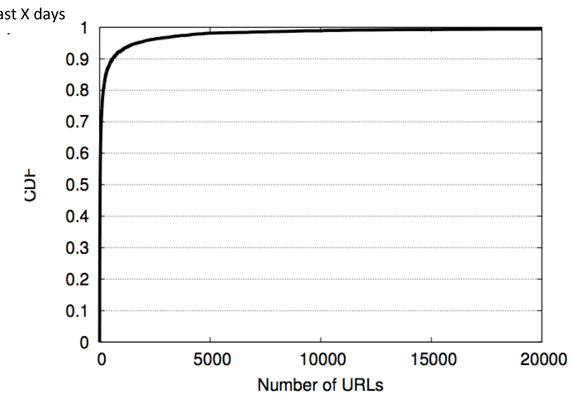
The Dataset consists of 150 million tweets collected from January to March 2014. All tweets contain PTs. In total 20.000 different trends were collected.

(1) Analysis of Popular Trends in Spam

Duration of a PT:
90% of PTs are active for only a single day.

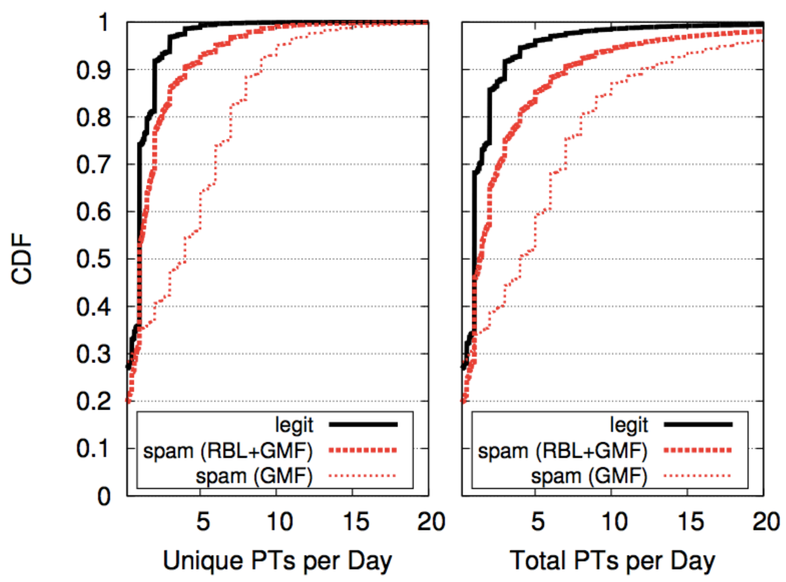


#URLs associated with a PT:
90% of PTs are associated with less than 1000 URLs



How much more PTs does Spam tweets contain?

- Compare with "classic" spam, or else RBL (Real-time Blackhole List) techniques.
- Percentage of users with more than 4 PTs:
  - Legit: 9%
  - GMF+RBL: 15%
  - GMF: 60% (!)

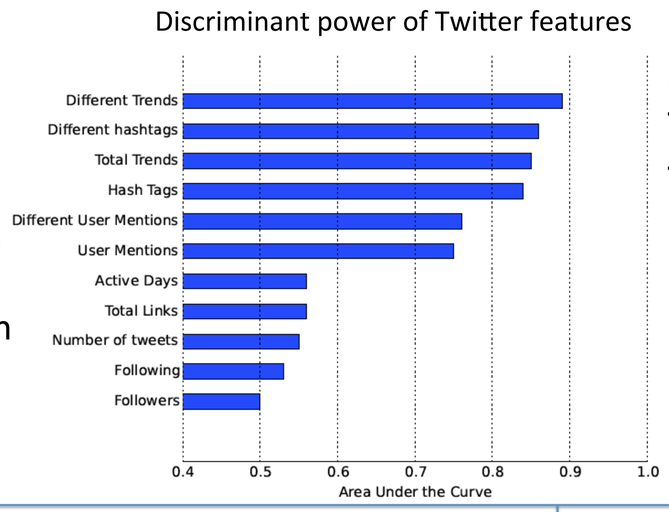


(2) Classify GMF spam vs legit tweets

Identify GMF spam based on Twitter features

- Decision Tree Regression classifier
- Repeated random sub-sampling validation
- TPR = True Positive Rate = TP/(TP+FN),
- FPR = False Positive Rate = FP/(FP+TN)
- User Classification: **TPR: 75.1%, FPR: 0.26%**
- Tweet Classification: **TPR: 81%, FPR: 0.18%**

Which features are the most discriminant?
Measure the AUC of the ROC curve by applying the classification schema for each one of the features. Most discriminant feature:
"Number of Different PTs"



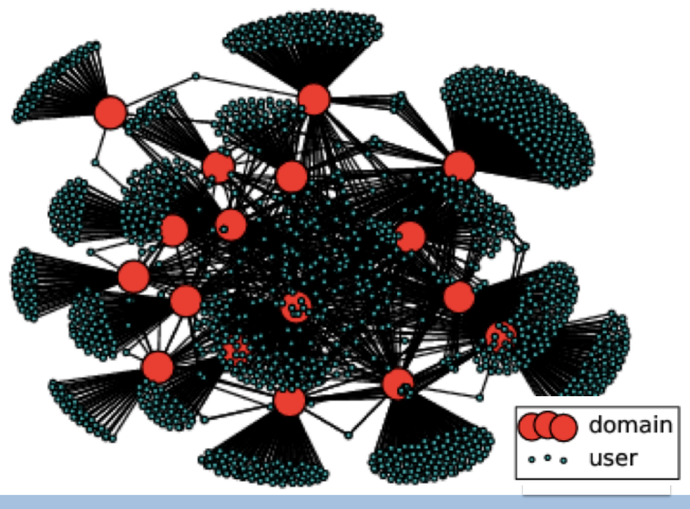
What is the most generic classification rule?
The rule on the root of the decision tree was:

if Unique PTs > 3.5 then Spam!
(TPR: 80%, FPR: 0.8%)

(3) Visualize GMF campaigns

How does a GMF campaign looks like?
Make a graph than connects spam users with GMF domains for a single day (10/1/2014).
17 GMF domains, only 2 IPs, 1.604 users.

Graph properties of spam campaigns
Graph properties can reveal the nature of spam campaigns and help build better classification methods.
For example: Average edge degree for spam=125, for legit=2.3



Future work

- Reveal more undetected spam techniques (link obfuscation)
- Develop filtering mechanisms based on graph properties
- Implement distributed detection on OSNs (gossip learning)

FORTH Foundation for Research & Technology - Hellas
MARIE CURIE ACTIONS
PEOPLE