中国科学院
CHINESE ACADEMY OF SCIENCES

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

Discovering the Fake Followers in the Micro-blogging via Machine Learning

Yi Shen Jianjun Yu

October 16, 2013

Computer Network Information Center, Chinese Academy of Sciences

Micro-blogging

twitter

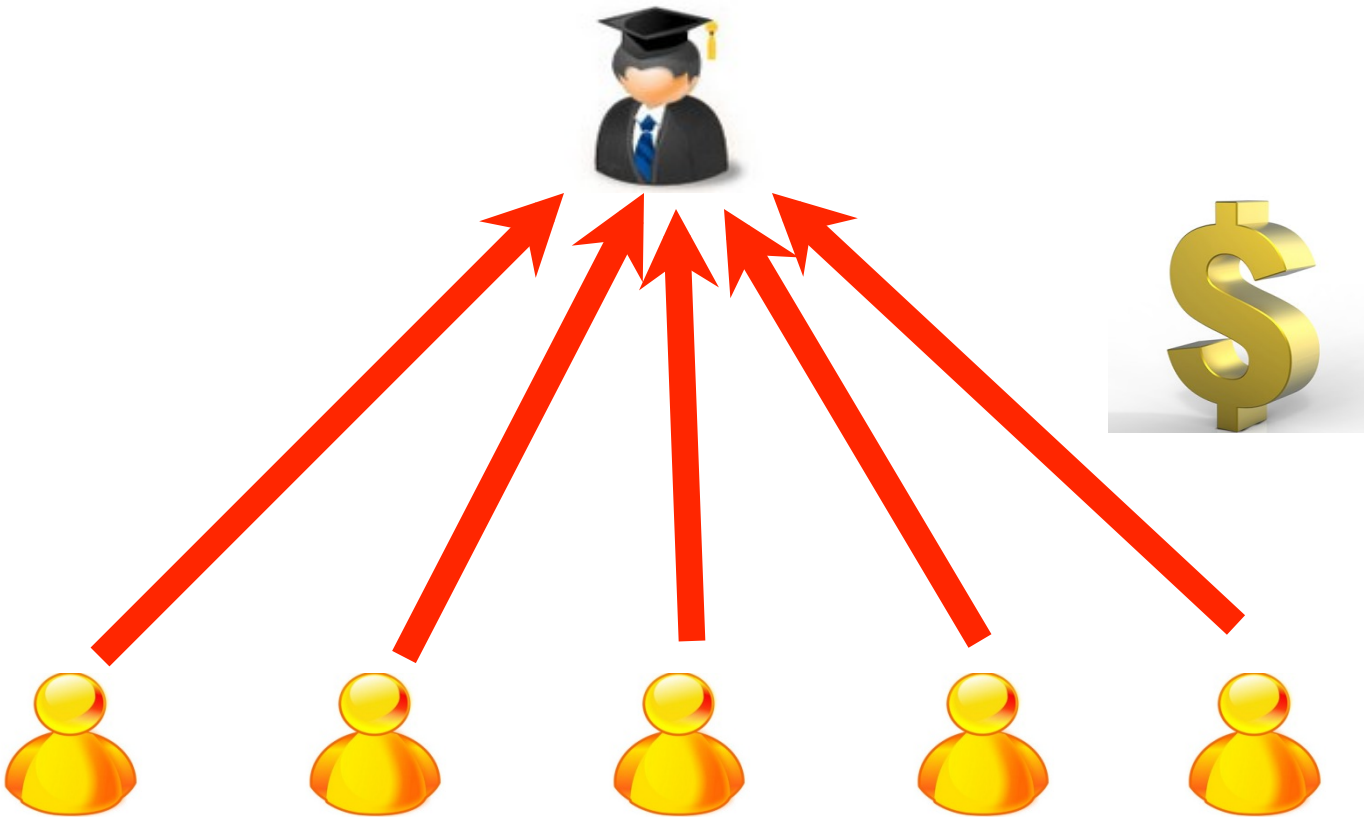
新浪微博
weibo.com

The Celebrities in Twitter

China Science & Technology Network

	Barack Obama	10,150 TWEETS	656,744 FOLLOWING	38,255,440 FOLLOWERS
	Justin Bieber	23,853 TWEETS	121,489 FOLLOWING	45,787,655 FOLLOWERS
	Kobe Bryant	626 TWEETS	844 FOLLOWING	3,599,653 FOLLOWERS
	Lady Gaga	3,535 TWEETS	134,762 FOLLOWING	40,249,744 FOLLOWERS

Purchasing Fake Followers



Fake followers Markets

Market Link	Price For 1K Followers
intertwitter.com	$9
solarank.com	$6.95
purchase-twitter-followers.net	$7.5
fakefollowerstwitter.com	$20

A Data Report

39% of @facebook followers are fake
34% of @ladygaga followers are fake
31% of @justinbieber followers are fake
32% of @katyperry followers are fake
32% of @espn followers are fake
33% of @britneyspears followers are fake
27% of @youtube followers are fake

http://www.socialsellingu.com/fake-twitter-profiles-infographic/

Troubles Caused by Fake Followers

- Noise for Social Network analysis
- Privacy and Security Problem
- Spam Problem

Method of Detection

- Binary Classification Problem
- Extract discriminative features
- Voting-SVM as the classifier

How to get ground-truth data?

- Purchase from different merchants
- Keep tracking them for a long period

The Features for Classification

- The Ratio of Followee Count and Follower Count (RFF)
- The Percentage of Bidirectional Friends (PBF)
- Average Repost Frequency of the Posts (ARF)
- Ratio of the Original Posts (ROP)
- Proportion of Nighttime Posts (PNP)
- Topic Diversity
- The standard deviation of post-count(σ_{post}).
- The general slope of post-count(g_{post}).
- The standard deviation of followee-count($\sigma_{followee}$)
- The decrease frequency of followee-count($DF_{followee}$).
- The standard deviation of follower-count ($\sigma_{follower}$).

Result

Accuracy	Precision	Recall	F1
98.1%	97.7%	96.6%	0.964

Thank you!