Link Farming in Twitter

Pawan Goyal

CSE, IITKGP

July 31, 2014

Reference

Saptarshi Ghosh, Bimal Viswanath, Farshad Kooti, Naveen Kumar Sharma, Korlam Gautam, Fabricio Benevenuto, Niloy Ganguly, and Krishna P. Gummadi. 2012. *Understanding and Combating Link Farming in the Twitter Social Network*. Proceedings of the 21st International World Wide Web Conference (WWW), Lyon, France.

Link Farming

Link Farming in Web

Websites exchange reciprocal links with other sites to improve ranking by search engines

Link Farming

Link Farming in Web

Websites exchange reciprocal links with other sites to improve ranking by search engines

Why Link Farming is an issue?

Search engines rank websites / webpages based on graph metrics such as Pagerank/HITS

- High in-degree helps to get high pagerank

Link Farming

Link Farming in Web

Websites exchange reciprocal links with other sites to improve ranking by search engines

Why Link Farming is an issue?

Search engines rank websites / webpages based on graph metrics such as Pagerank/HITS

- High in-degree helps to get high pagerank

A form of spam

Heavily penalized by search engines

Link Farming in Twitter

Twitter as a Web within the Web

- Vast amount of information and real-time news
- Twitter search becoming more and more common

Link Farming in Twitter

Twitter as a Web within the Web

- Vast amount of information and real-time news
- Twitter search becoming more and more common
- Search engines rank users by follower-rank, Pagerank to decide whose tweet to return as search results

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{j,k} + \frac{1-\mu}{|S|}$$

Link Farming in Twitter

Twitter as a Web within the Web

- Vast amount of information and real-time news
- Twitter search becoming more and more common
- Search engines rank users by follower-rank, Pagerank to decide whose tweet to return as search results

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{j,k} + \frac{1-\mu}{|S|}$$

- High indegree (no. of followers) is seen as a metric of influence

Link Farming in Twitter

Twitter as a Web within the Web

- Vast amount of information and real-time news
- Twitter search becoming more and more common
- Search engines rank users by follower-rank, Pagerank to decide whose tweet to return as search results

$$I_j = \mu \cdot \sum_{\forall k \neq j} I_k \cdot \tilde{M}_{j,k} + \frac{1-\mu}{|S|}$$

- High indegree (no. of followers) is seen as a metric of influence

Link Farming in Twitter

Spammers follow other users and attempt to get them to follow back

Link farming in Web and Twitter

Motivation is similar

Higher indegree will give better ranks in search results

Link farming in Web and Twitter

Motivation is similar

Higher indegree will give better ranks in search results

Who engages in link farming?

- Web - spammers
- Twitter - spammers,

Link farming in Web and Twitter

Motivation is similar

Higher indegree will give better ranks in search results

Who engages in link farming?

- Web - spammers
- Twitter - spammers, many legitimate, popular users

Link farming in Web and Twitter

Motivation is similar

Higher indegree will give better ranks in search results

Who engages in link farming?

- Web - spammers
- Twitter - spammers, many legitimate, popular users

Additional factors in Twitter

'Following back' considered as a social ettiquette

Link farmers in Twitter

Idea: start with spammers

Study how spammers acquire social links

Link farmers in Twitter

Idea: start with spammers

Study how spammers acquire social links

Large amounts of spam in Twitter

- Spam-URLs get much higher clickthrough rates than spam-URLs in email
- Spammers are successfully acquiring social links and social influence

Dataset

Twitter dataset collected at MPI-SWS, Germany

- Complete snapshot of Twitter as of August 2009
- 54 million users, 1.9 billion social links

Dataset

Twitter dataset collected at MPI-SWS, Germany

- Complete snapshot of Twitter as of August 2009
- 54 million users, 1.9 billion social links

Identifying spammers

Attempt to crawl user's profile page - if the user is suspended, crawl would lead to `http://twitter.com/suspended`

Dataset

Twitter dataset collected at MPI-SWS, Germany

- Complete snapshot of Twitter as of August 2009
- 54 million users, 1.9 billion social links

Identifying spammers

Attempt to crawl user's profile page - if the user is suspended, crawl would lead to `http://twitter.com/suspended`

- 379,340 accounts suspended during Aug 2009 - Feb 2011

Dataset

Twitter dataset collected at MPI-SWS, Germany

- Complete snapshot of Twitter as of August 2009
- 54 million users, 1.9 billion social links

Identifying spammers

Attempt to crawl user's profile page - if the user is suspended, crawl would lead to `http://twitter.com/suspended`

- 379,340 accounts suspended during Aug 2009 - Feb 2011
- Suspension - either due ot spam-activity or long inactivity

Dataset

Twitter dataset collected at MPI-SWS, Germany

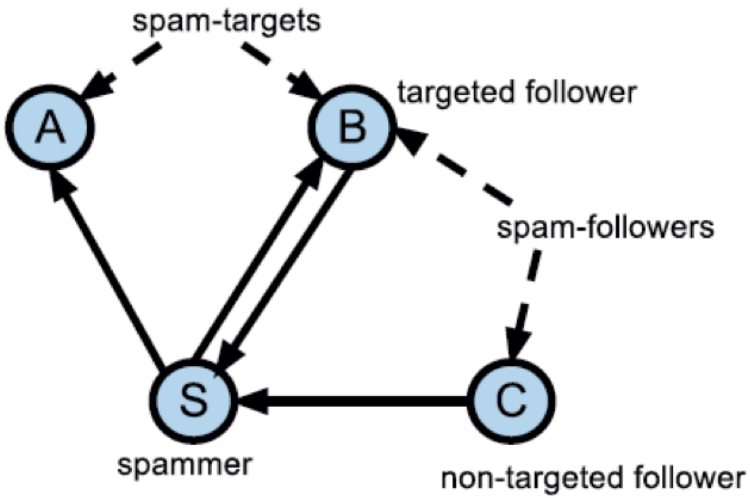
- Complete snapshot of Twitter as of August 2009
- 54 million users, 1.9 billion social links

Identifying spammers

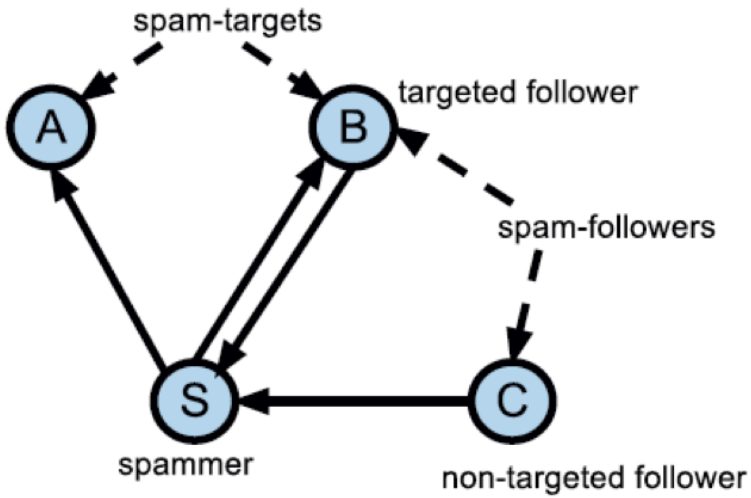
Attempt to crawl user's profile page - if the user is suspended, crawl would lead to `http://twitter.com/suspended`

- 379,340 accounts suspended during Aug 2009 - Feb 2011
- Suspension - either due ot spam-activity or long inactivity
- *41,352 suspended accounts posted at least one blacklisted URL shortened by bit.ly or tinyurl*

Terminology for spammers' links

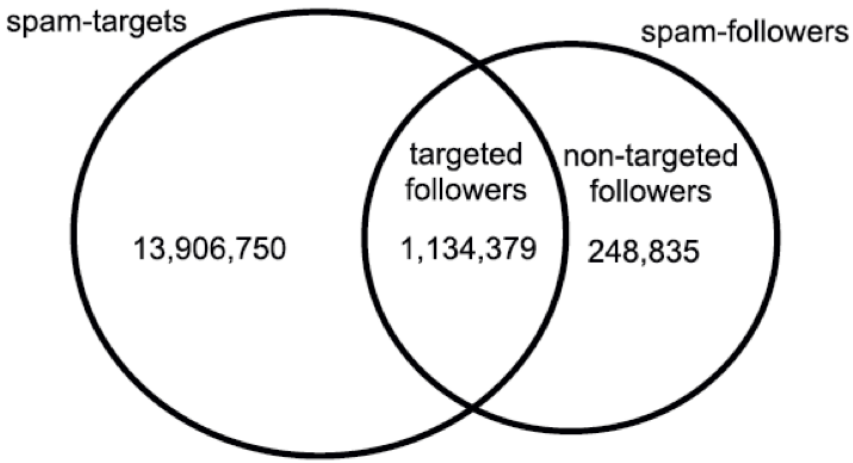


Terminology for spammers' links



- Spam-targets: users followed by spammers
- Spam-followers: users who follow spammers
- Targeted followers: spam-target as well as spam-follower
- Non-targeted followers: follow spammers without being targeted

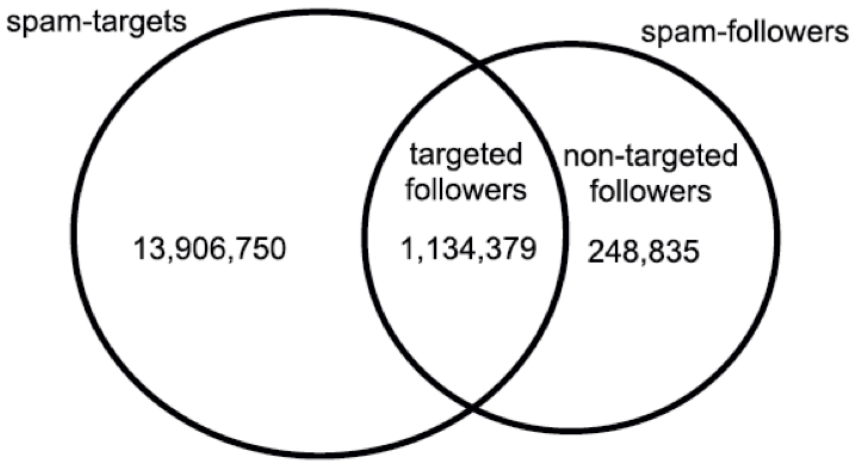
Link farming by spammers



Spammers farm links at large scale

Over 13 million users (27% of total) targeted by 41,352 spammers (0.08% of total)

Link farming by spammers



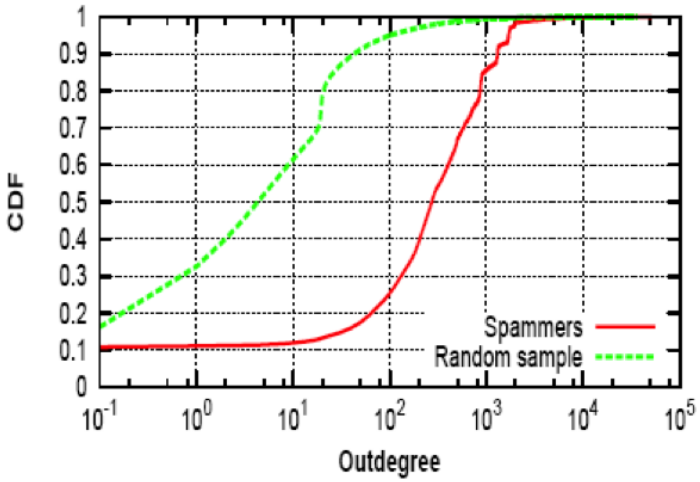
Spammers farm links at large scale

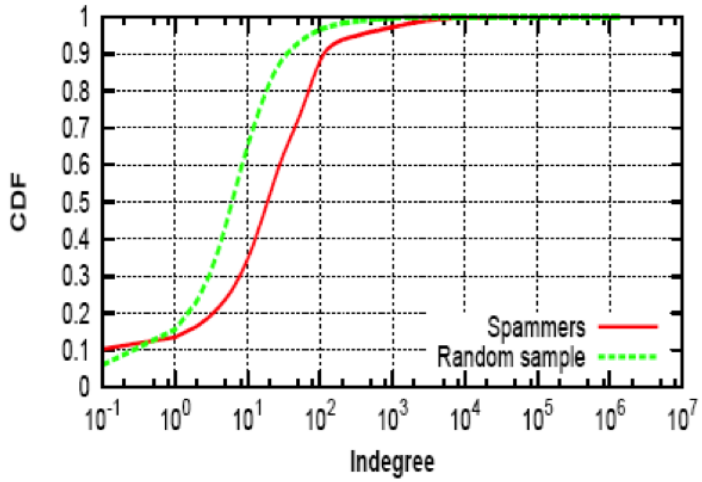
Over 13 million users (27% of total) targeted by 41,352 spammers (0.08% of total)

1.3 million spam-followers

82% are targeted $\rightarrow$ spammers get most links by reciprocation

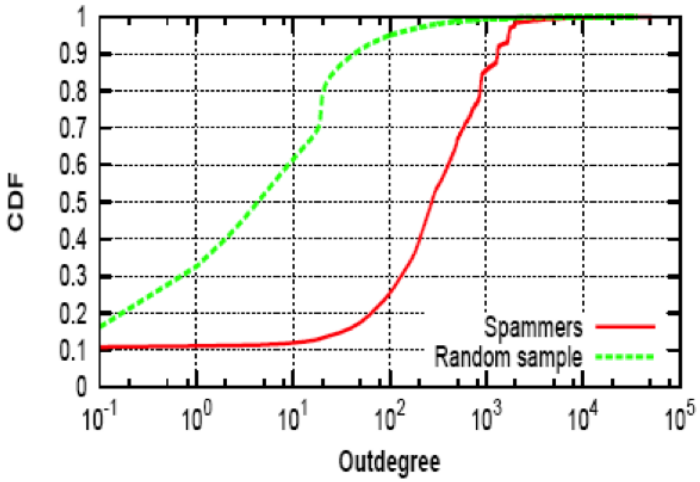
Link farming makes spammers influential

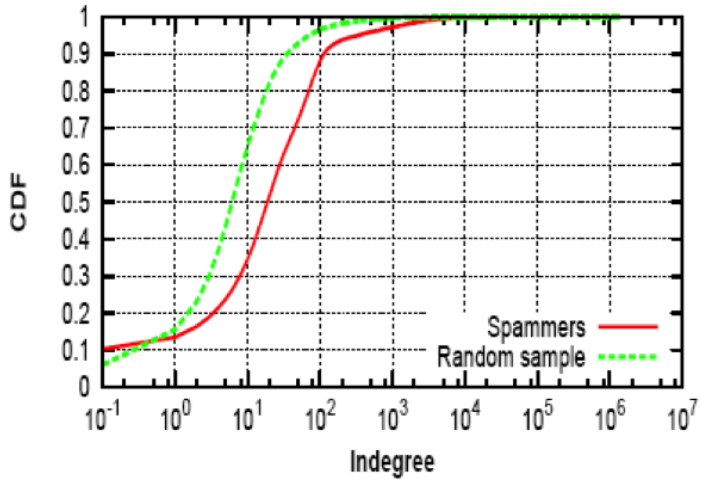




- Spammers get more followers than an average Twitter user
- Some spammers acquire very high Pageranks :

Link farming makes spammers influential





- Spammers get more followers than an average Twitter user
- Some spammers acquire very high Pageranks : 304 with top 100,000 (0.18% of all users)

Who are the spam-followers?

Non-targeted spam-followers

- Mostly sybils / hired helps of spammers
- Most have now been suspended by Twitter (9,725 among top 10,000, having links to spammers)

Who are the spam-followers?

Non-targeted spam-followers

- Mostly sybils / hired helps of spammers
- Most have now been suspended by Twitter (9,725 among top 10,000, having links to spammers)

Targeted spam-followers

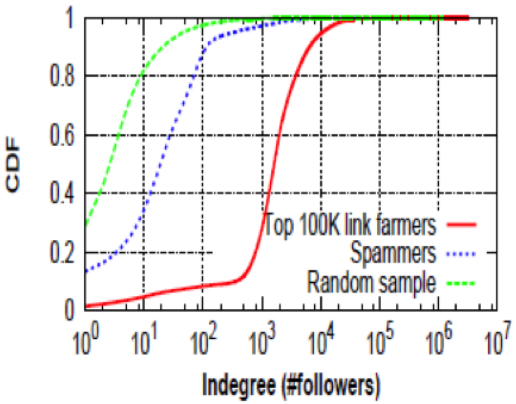
- Ranked on the basis of number of links to spammers
- 60% of the follow-links acquired by spammers come from the top 100,000 targeted followers

Who are the top link-farmers?

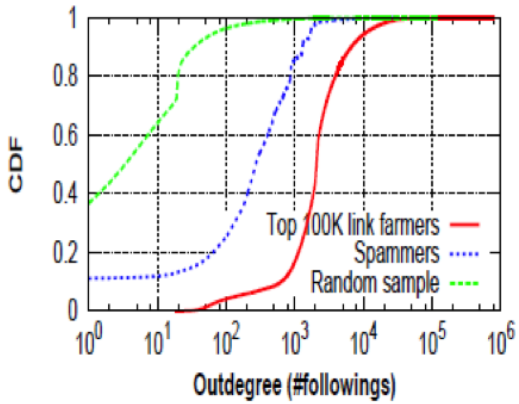
- Analyzed the status of the top 100,000 link farmers (July, 2011)
- 76% still exist and have not been suspended by Twitter
- 235 verified as real, well-known users
- much higher indegree as well as outdegree compared to spammers
- Most of their tweets contain valid URLs

Who are the top link-farmers?

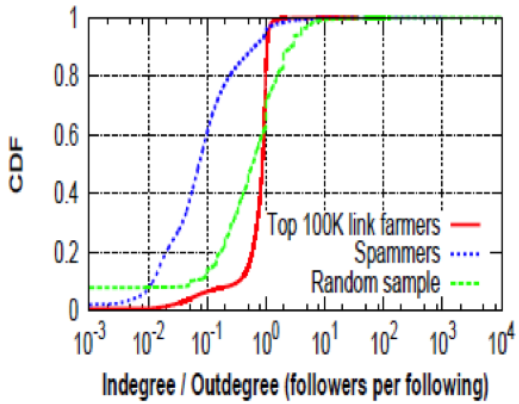
- Analyzed the status of the top 100,000 link farmers (July, 2011)
- 76% still exist and have not been suspended by Twitter
- 235 verified as real, well-known users
- much higher indegree as well as outdegree compared to spammers
- Most of their tweets contain valid URLs



(a) Indegree



(b) Outdegree



(c) Indegree/outdegree ratio

Who are the top link-farmers?

Top 5 link farmers according to	
#links to spammers	**Pagerank**
Larry Wentz: Internet, Affiliate Marketing	Barack Obama: campaign staff
Judy Rey Wasserman: Artist, founder	Britney Spears: It's Britney
Chris Latko: Interested in tech. Will follow back	NPR Politics: Political coverage and conversation
Paul Merriwether: helping others, let's talk soon	UK Prime Minister: PM's office
Aaron Lee: Social Media Manager	JetBlue Airways: Follow us and let us help

Who are the top link-farmers?

Highly influential users

Rank within top 5% as per Pagerank, follower-rank, retweet-rank

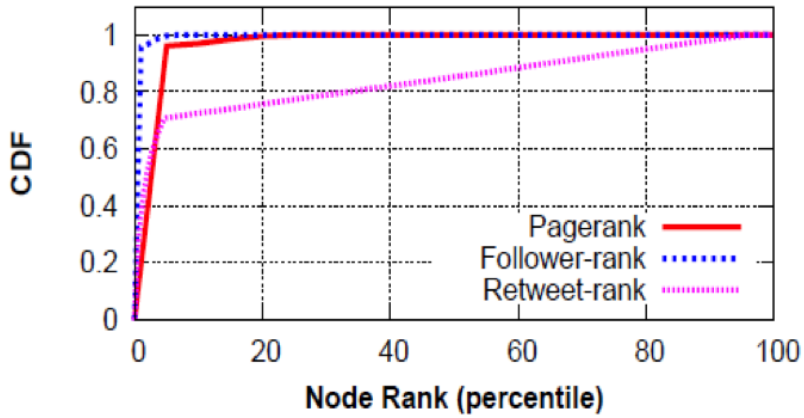
Who are the top link-farmers?

Highly influential users

Rank within top 5% as per Pagerank, follower-rank, retweet-rank

Mostly social marketers, enterpreneurs, ...

- Want to promote some online business/website
- Heavily interconnect with each other - density 0.018 (10^{-7} for the whole graph)
- Aim: to acquire social capital



Combating the problem

Not practical for Twitter to suspend / blacklist top link-farmers

Combating the problem

Not practical for Twitter to suspend / blacklist top link-farmers

Solution

- Strategy to disincentivize users from following / reciprocating to unknown people
- Penalize users for following spammers

Combating the problem

Not practical for Twitter to suspend / blacklist top link-farmers

Solution

- Strategy to disincentivize users from following / reciprocating to unknown people
- Penalize users for following spammers

Collusionrank: inverse of pagerank

- Negatively bias a small set of known spammers
- Propagate negative scores from spammers to spam-followers

Collusionrank

Algorithm 1 Collusionrank

Input: network, G; set of known spammers, S; decay factor for biased Pagerank, α

Output: Collusionrank scores, c

initialize score vector d for all nodes n in G

$$d(n) \leftarrow \begin{cases} \frac{-1}{|S|} & \text{if} n \in S \\ 0 & \text{otherwise} \end{cases}$$

/* compute Collusionrank scores */

$c \leftarrow d$

while c not converged **do**

  for all nodes n in G **do**

$$tmp \leftarrow \sum_{nbr \in followings(n)} \frac{c(nbr)}{|followers(nbr)|}$$

$$c(n) \leftarrow \alpha * tmp + (1 - \alpha) * d(n)$$

  end for

  insert leaked scores uniformly across all nodes such that $\sum_n c(n) = -1$

end while

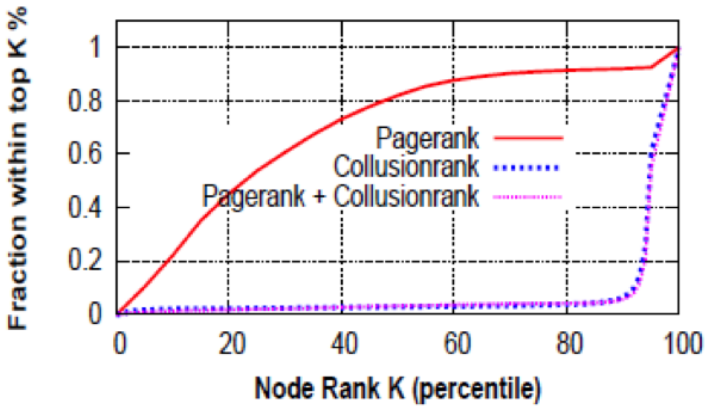
return c

Pagerank+Collusionrank

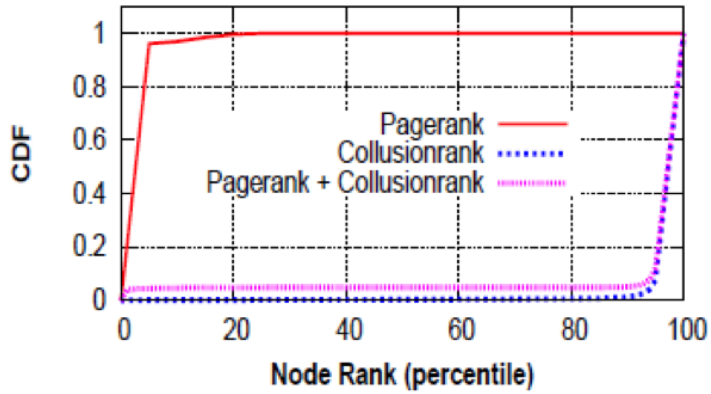
- Computed Collusionrank considering 600 known spammers
- Rank users by Pagerank + Collusionrank

Pagerank+Collusionrank

- Computed Collusionrank considering 600 known spammers
- Rank users by Pagerank + Collusionrank
  $\rightarrow$ Effectively filters out spammers and link-farmers (top spam-followers) from top ranks



(a) Rankings of all 41,352 spammers



(b) Rankings of Top 100,000 capitalists