# Privacy and Security in Online Social Media

Course on NPTEL

NOC21-CS28

Week 6.2

Ponnurangam Kumaraguru ("PK")
Full Professor
ACM Distinguished Speaker
fb/ponnurangam.kumaraguru, @ponguru

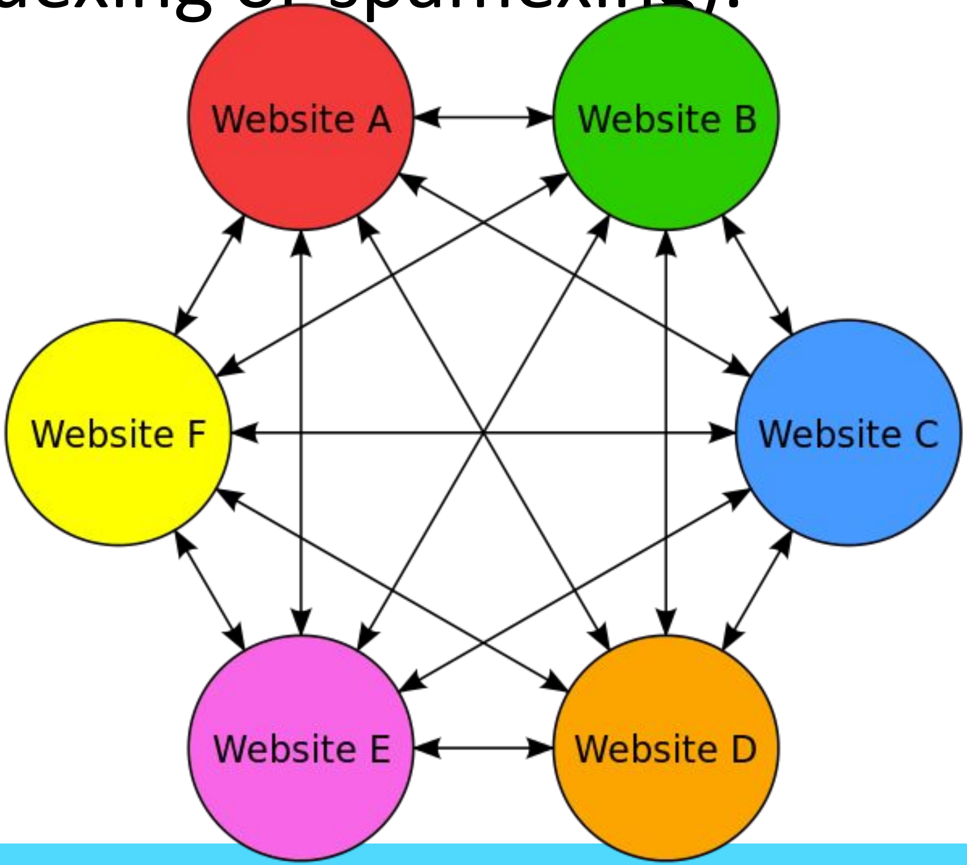
# Link farming

- Search engines rank websites / webpages based on graph metrics such as Pagerank
  - High in-degree helps to get high Pagerank

- Link farming in Web
  - Websites exchange reciprocal links with other sites to improve ranking by search engines

# Link farming

- A link farm is a form of spamming the index of a search engine (sometimes called spamdexing or spamexing).



# Why link farming in Twitter?

- Twitter has become a Web within the Web
  - Vast amounts of information and real-time news
  - Twitter search becoming more and more common
  - Search engines rank users by follower-rank, Pagerank to decide whose tweets to return as search results
  - High indegree (#followers) seen as a metric of influence
  - Klout score influenced by Twitter indegree

- Link farming in Twitter
  - Spammers follow other users and attempt to get them to follow back (Reciprocity)

# Link farming in Web & Twitter similar?

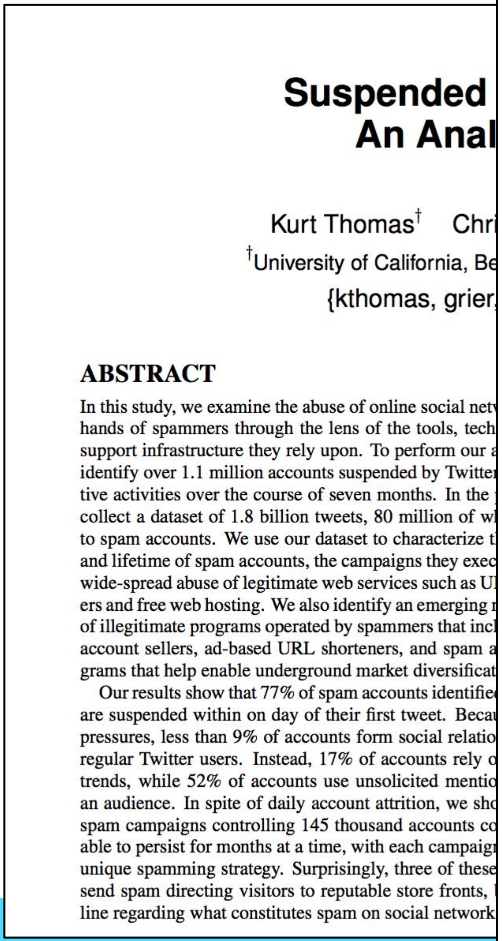
- Motivation is similar
  - Higher indegree will give better ranks in search results
- Who engages in link farming?
  - Web – spammers
  - Twitter – spammers + many legitimate, popular users !!!
- Additional factors in Twitter
  - ‘Following back’ considered a social etiquette

# Is link farming in Twitter spam at all?

- Your reactions?

# Spam in Twitter

- “five spam campaigns controlling 145 thousand accounts combined are able to persist for months at a time, with each campaign enacting a unique spamming strategy.”

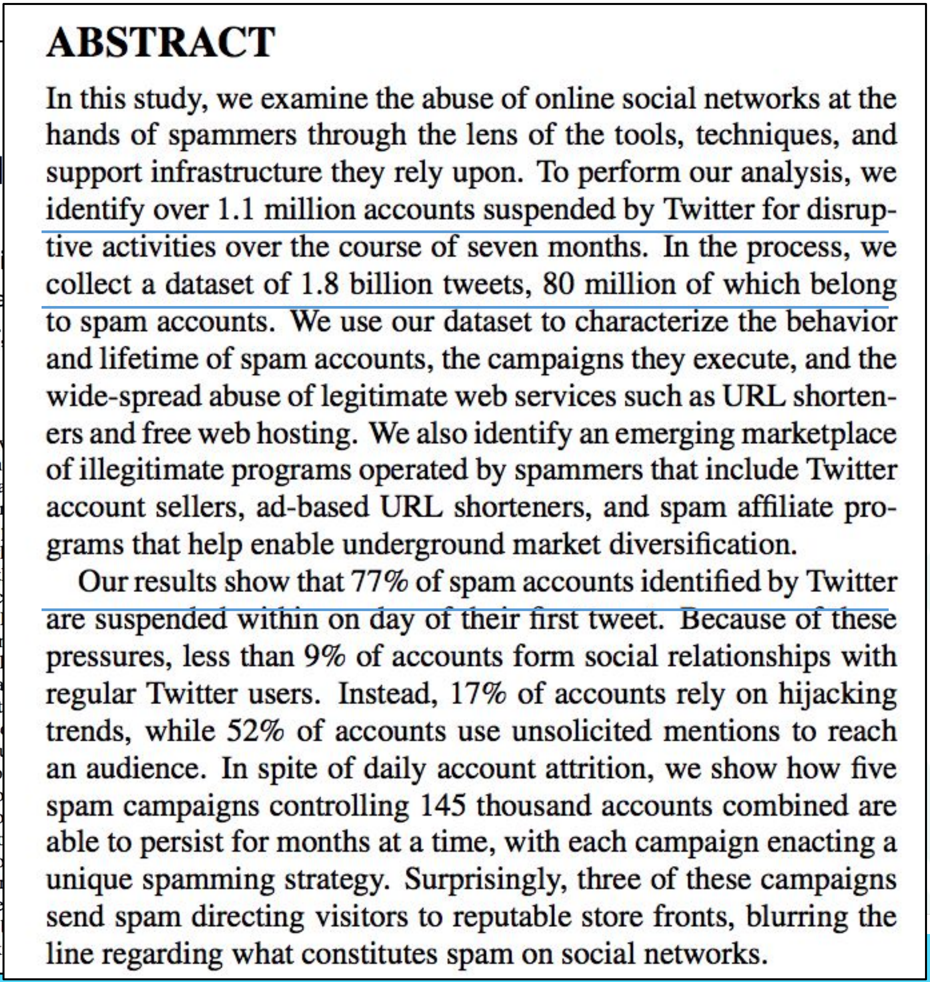


**Suspended**
**An Anal**

Kurt Thomas[†] Chri
[†]University of California, Be
{kthomas, grier,

ABSTRACT

In this study, we examine the abuse of online social netw
hands of spammers through the lens of the tools, tech
support infrastructure they rely upon. To perform our a
identify over 1.1 million accounts suspended by Twitter
tive activities over the course of seven months. In the
collect a dataset of 1.8 billion tweets, 80 million of wh
to spam accounts. We use our dataset to characterize t
and lifetime of spam accounts, the campaigns they exec
wide-spread abuse of legitimate web services such as U
ers and free web hosting. We also identify an emerging
of illegitimate programs operated by spammers that incl
account sellers, ad-based URL shorteners, and spam a
grams that help enable underground market diversificat
Our results show that 77% of spam accounts identifie
are suspended within on day of their first tweet. Becau
pressures, less than 9% of accounts form social relatio
regular Twitter users. Instead, 17% of accounts rely o
trends, while 52% of accounts use unsolicited mentio
an audience. In spite of daily account attrition, we sho
spam campaigns controlling 145 thousand accounts co
able to persist for months at a time, with each campaig
unique spamming strategy. Surprisingly, three of these
send spam directing visitors to reputable store fronts,
line regarding what constitutes spam on social network



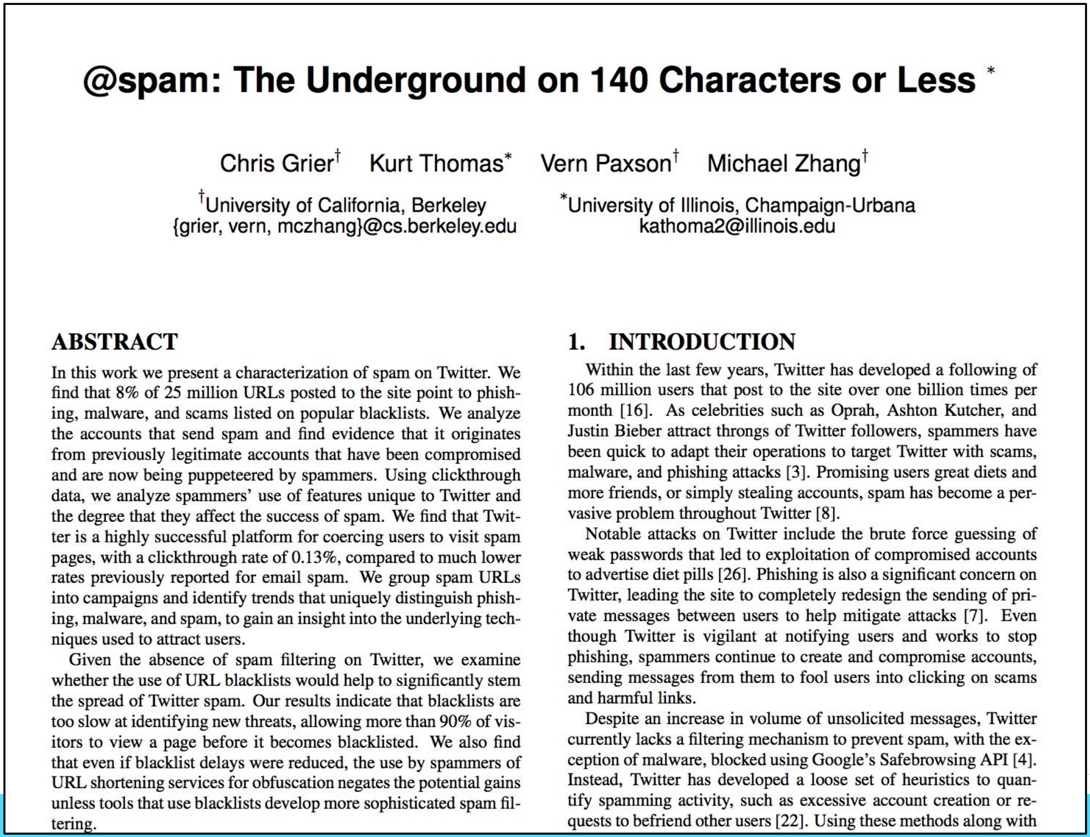
## ABSTRACT

In this study, we examine the abuse of online social networks at the hands of spammers through the lens of the tools, techniques, and support infrastructure they rely upon. To perform our analysis, we identify over 1.1 million accounts suspended by Twitter for disruptive activities over the course of seven months. In the process, we collect a dataset of 1.8 billion tweets, 80 million of which belong to spam accounts. We use our dataset to characterize the behavior and lifetime of spam accounts, the campaigns they execute, and the wide-spread abuse of legitimate web services such as URL shorteners and free web hosting. We also identify an emerging marketplace of illegitimate programs operated by spammers that include Twitter account sellers, ad-based URL shorteners, and spam affiliate programs that help enable underground market diversification.

Our results show that 77% of spam accounts identified by Twitter are suspended within on day of their first tweet. Because of these pressures, less than 9% of accounts form social relationships with regular Twitter users. Instead, 17% of accounts rely on hijacking trends, while 52% of accounts use unsolicited mentions to reach an audience. In spite of daily account attrition, we show how five spam campaigns controlling 145 thousand accounts combined are able to persist for months at a time, with each campaign enacting a unique spamming strategy. Surprisingly, three of these campaigns send spam directing visitors to reputable store fronts, blurring the line regarding what constitutes spam on social networks.

# Spam in Twitter

- “We find that 8% of 25 million URLs posted to the site point to phishing, malware, and scams listed on popular blacklists.”
- “We find that Twitter is a highly successful platform for coercing users to visit spam pages, with a clickthrough rate of 0.13%, compared to much lower rates previously reported for email spam”



## @spam: The Underground on 140 Characters or Less [*]

Chris Grier[†] Kurt Thomas[*] Vern Paxson[†] Michael Zhang[†]

[†]University of California, Berkeley
{grier, vern, mczhang}@cs.berkeley.edu

[*]University of Illinois, Champaign-Urbana
kathoma2@illinois.edu

### ABSTRACT

In this work we present a characterization of spam on Twitter. We find that 8% of 25 million URLs posted to the site point to phishing, malware, and scams listed on popular blacklists. We analyze the accounts that send spam and find evidence that it originates from previously legitimate accounts that have been compromised and are now being puppeteered by spammers. Using clickthrough data, we analyze spammers' use of features unique to Twitter and the degree that they affect the success of spam. We find that Twitter is a highly successful platform for coercing users to visit spam pages, with a clickthrough rate of 0.13%, compared to much lower rates previously reported for email spam. We group spam URLs into campaigns and identify trends that uniquely distinguish phishing, malware, and spam, to gain an insight into the underlying techniques used to attract users.

Given the absence of spam filtering on Twitter, we examine whether the use of URL blacklists would help to significantly stem the spread of Twitter spam. Our results indicate that blacklists are too slow at identifying new threats, allowing more than 90% of visitors to view a page before it becomes blacklisted. We also find that even if blacklist delays were reduced, the use by spammers of URL shortening services for obfuscation negates the potential gains unless tools that use blacklists develop more sophisticated spam filtering.

### 1. INTRODUCTION

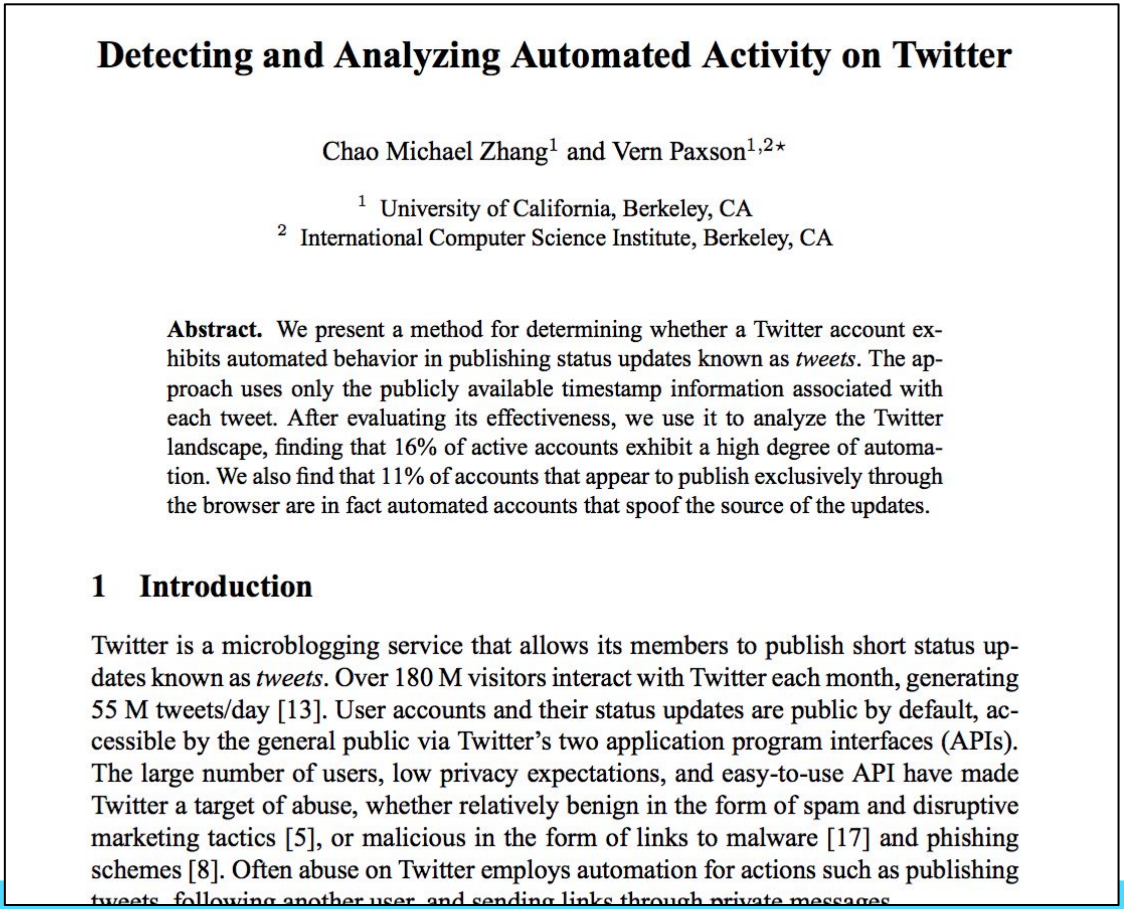
Within the last few years, Twitter has developed a following of 106 million users that post to the site over one billion times per month [16]. As celebrities such as Oprah, Ashton Kutcher, and Justin Bieber attract throngs of Twitter followers, spammers have been quick to adapt their operations to target Twitter with scams, malware, and phishing attacks [3]. Promising users great diets and more friends, or simply stealing accounts, spam has become a pervasive problem throughout Twitter [8].

Notable attacks on Twitter include the brute force guessing of weak passwords that led to exploitation of compromised accounts to advertise diet pills [26]. Phishing is also a significant concern on Twitter, leading the site to completely redesign the sending of private messages between users to help mitigate attacks [7]. Even though Twitter is vigilant at notifying users and works to stop phishing, spammers continue to create and compromise accounts, sending messages from them to fool users into clicking on scams and harmful links.

Despite an increase in volume of unsolicited messages, Twitter currently lacks a filtering mechanism to prevent spam, with the exception of malware, blocked using Google's Safebrowsing API [4]. Instead, Twitter has developed a loose set of heuristics to quantify spamming activity, such as excessive account creation or requests to befriend other users [22]. Using these methods along with

# Spam in Twitter

- “finding that 16% of active accounts exhibit a high degree of automation.”
- “find that 11% of accounts that appear to publish exclusively through the browser are in fact automated accounts that spoof the source of the updates.”



## Detecting and Analyzing Automated Activity on Twitter

Chao Michael Zhang[1] and Vern Paxson[1,2★]

[1] University of California, Berkeley, CA
[2] International Computer Science Institute, Berkeley, CA

**Abstract.** We present a method for determining whether a Twitter account exhibits automated behavior in publishing status updates known as *tweets*. The approach uses only the publicly available timestamp information associated with each tweet. After evaluating its effectiveness, we use it to analyze the Twitter landscape, finding that 16% of active accounts exhibit a high degree of automation. We also find that 11% of accounts that appear to publish exclusively through the browser are in fact automated accounts that spoof the source of the updates.

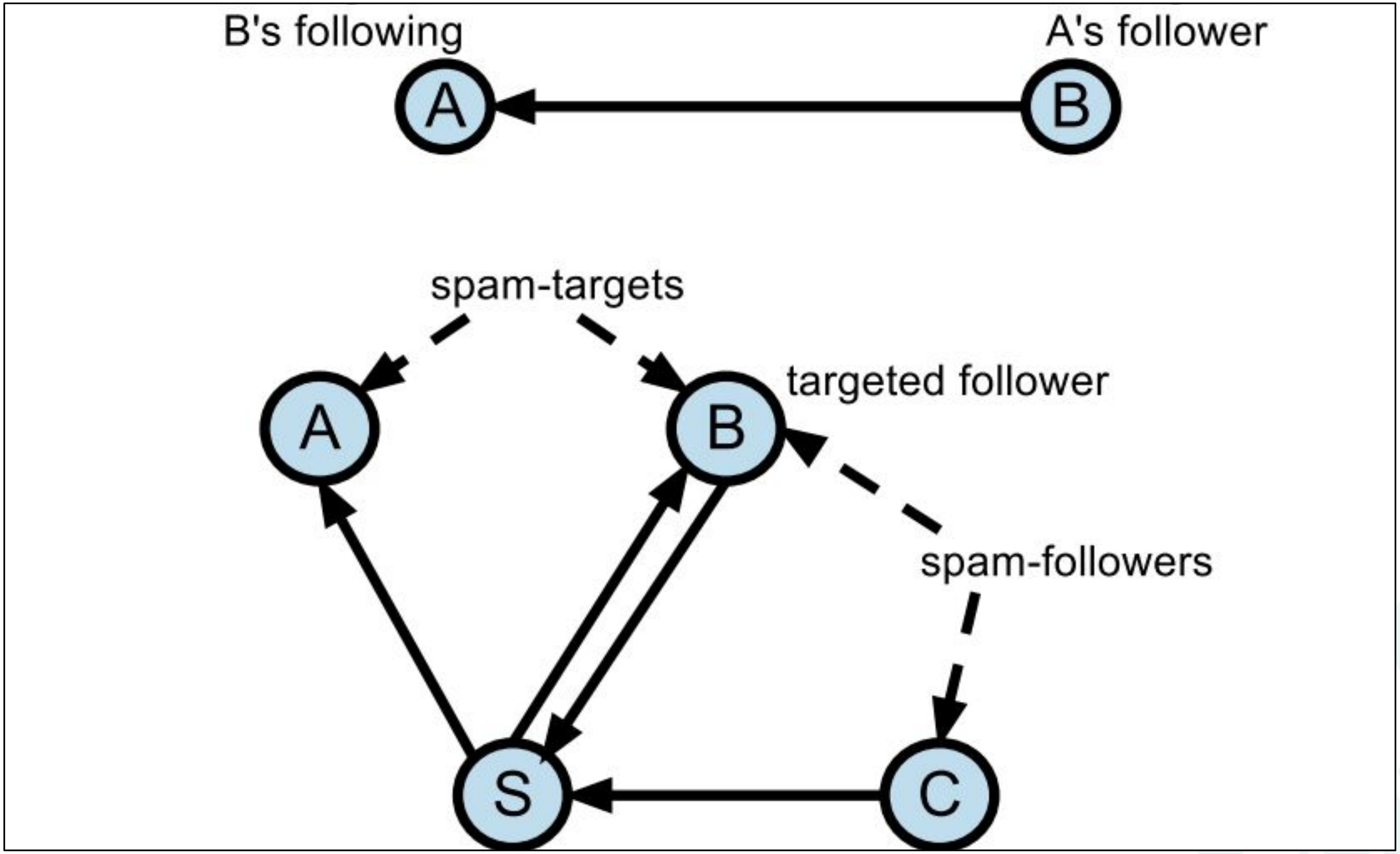
### 1 Introduction

Twitter is a microblogging service that allows its members to publish short status updates known as *tweets*. Over 180 M visitors interact with Twitter each month, generating 55 M tweets/day [13]. User accounts and their status updates are public by default, accessible by the general public via Twitter’s two application program interfaces (APIs). The large number of users, low privacy expectations, and easy-to-use API have made Twitter a target of abuse, whether relatively benign in the form of spam and disruptive marketing tactics [5], or malicious in the form of links to malware [17] and phishing schemes [8]. Often abuse on Twitter employs automation for actions such as publishing tweets, following another user, and sending links through private messages.

# Dataset

- Complete snapshot of Twitter, 2009
- 54 million users, 1.9 billion links! Largest dataset!

# Nodes

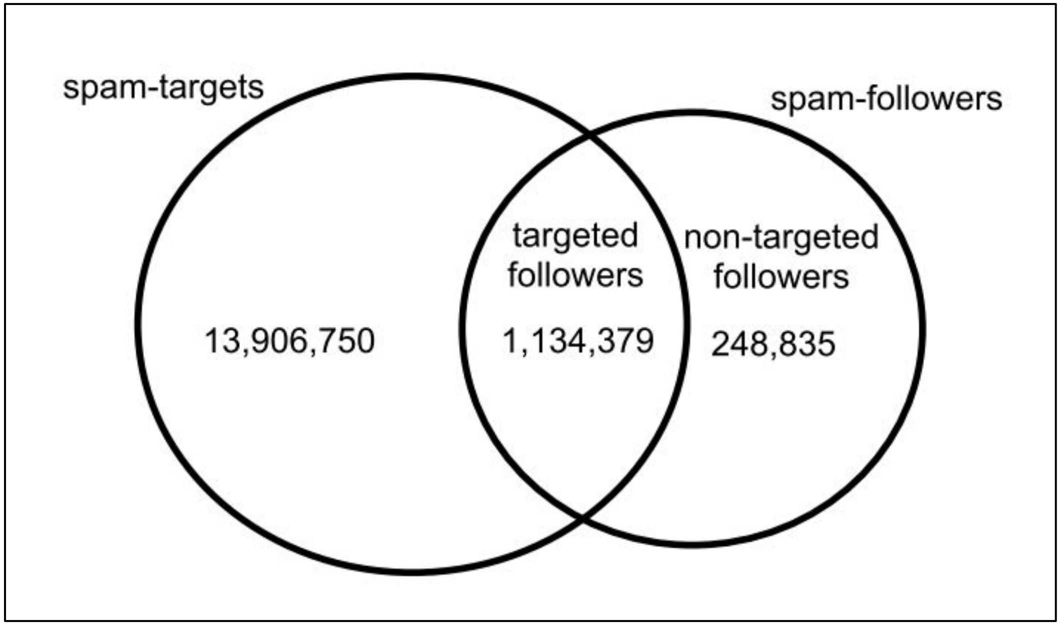

B's following
A's follower
A
B
spam-targets
targeted follower
A
B
spam-followers
S
C

# Spammers

- 379,340 accounts has been suspended in the interval, Aug 09 – Feb 11
  - Spam-activity or long inactivity
- 41,352 suspended accounts posted at least one blacklisted URL shortened by bitly, tinyurl

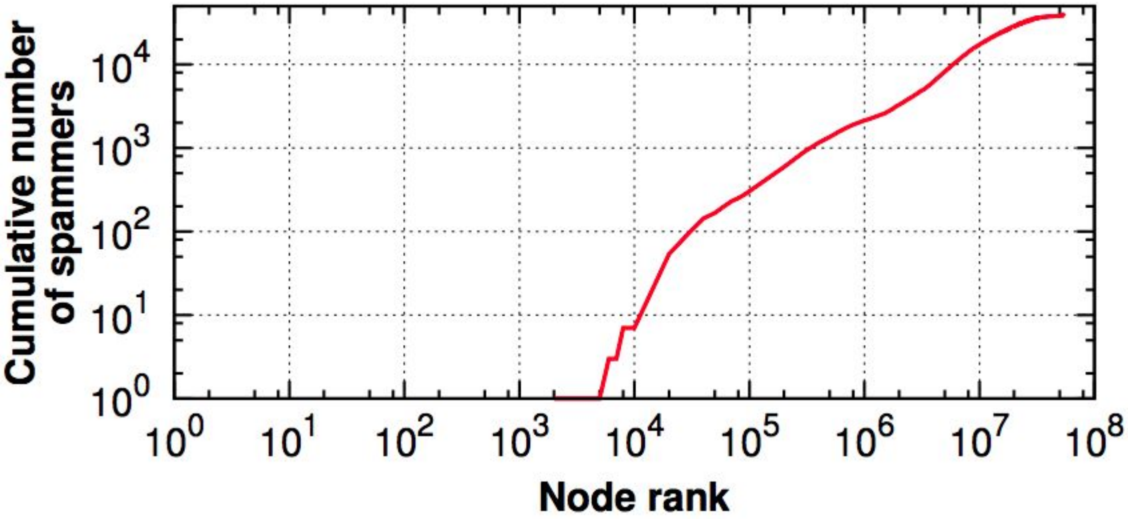
# Spammers

- # of spam-targets, spam-followers, their overlap
- 82% of spam followers overlap with the spam-targets



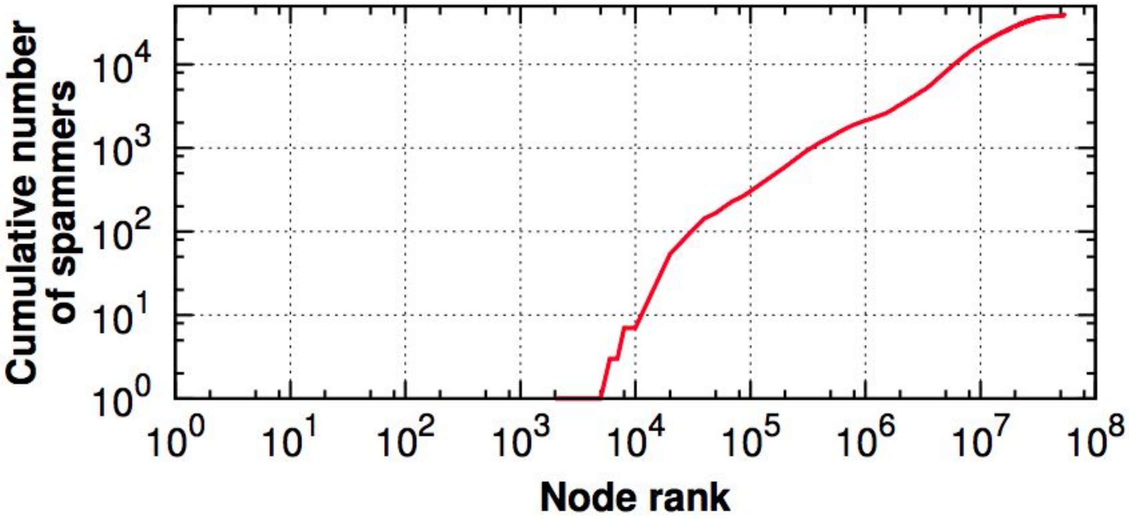
# Spammers

- # of spammers who rank within the top K according to pagerank



# Spammers

- # of spammers who rank within the top K according to pagerank
- 7 spammers rank within 10,000, 304 within 100,000 and 2,131 within 1million



# Thank you

pk@iiitd.ac.in
precog.iiitd.edu.in
fb/ponnurangam.kumaraguru