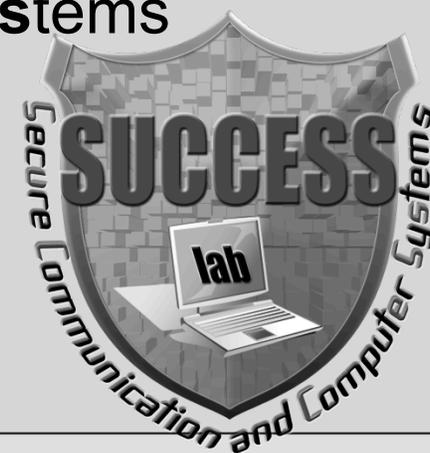


# Machine Learning Meets Social Networking Security: Detecting and Analyzing Malicious Social Networks for Fun and Profit

Guofei Gu

Secure Communication and Computer Systems  
(SUCCESS) Lab

Texas A&M University



## Credit

- Chao Yang, Robert Harkreader, Guofei Gu. "Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers." In Proceedings of the 14th International Symposium on Recent Advances in Intrusion Detection (RAID 2011), Menlo Park, California, September 2011
- Chao Yang, Robert Harkreader, Jialong Zhang, Suengwon Shin, and Guofei Gu. "Analyzing Spammers' Social Networks For Fun and Profit -- A Case Study of Cyber Criminal Ecosystem on Twitter." In Proceedings of the 21st International World Wide Web Conference (WWW'12), Lyon, France, April 2012

# Roadmap Today

---

- **Background**
  - Detecting Malicious OSN Identities
  - Analyzing Malicious Social Networks
  - Conclusion
-

# Introduction: OSNs are Popular



Follow your interests

Instant updates from your friends, industry experts, favorite celebrities, and what's happening around the world.





TEXAS A&M  
UNIVERSITY

# Background: OSNs are Suffering



 **Claimed my free iPhone** today, so happy lol... If anyone else wants one go here <http://tinyurl.com/36xjm3s>  
about a minute ago via Email

# Backgrounds: Attacks on Twitter

Twitter phishing hack hits BBC, PCC ...

and Gu  
and ba



Detect and suspend malicious OSN accounts individually



UKP  
here <http://tr.im/PPMS>  
about 10 hours ago

Alert: W32 Koobface Worm now on

A direct mess



Understand how criminal accounts survive and work on Twitter

circulating in Twitter.

Tweet

Like

Koobface has been spreading in the wild throughout the month of June targeting Facebook & MySpace accounts.



Analyze OSN criminal accounts' social relationships and ecosystem

the worm to his or her friends targeting more social networking w  
Tagged, Netlog and most recently, Twitter.

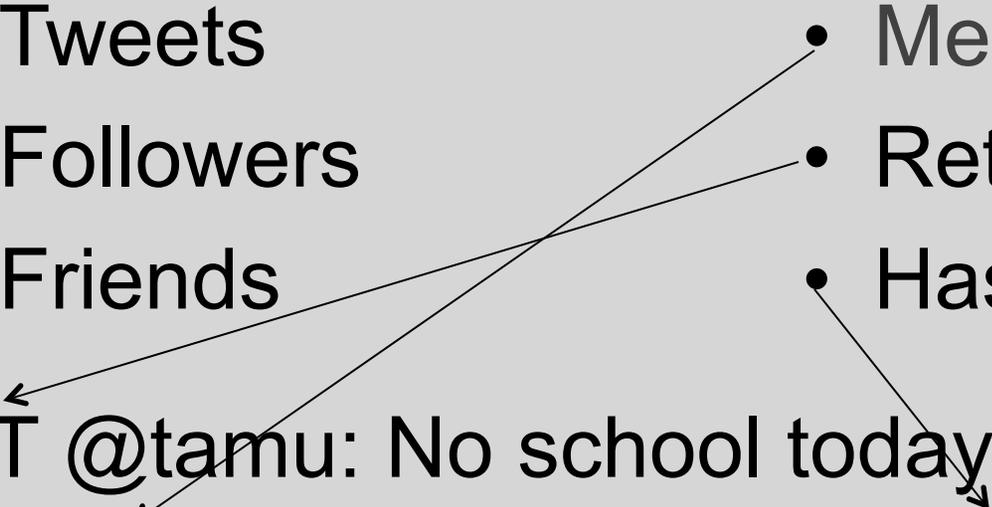
nts detected



# Twitter ABC

- What is Twitter?
  - Social media site
  - Informal information sharing
  - Messages limited to 140 characters
- Tweets
- Followers
- Friends
- Mentions
- Retweets
- Hashtags

RT @tamu: No school today!! U can thank @dustin. Go watch some #aggiefootball



# Introduction: Typical Behaviors of Spam Accounts

Follow Many Accounts



Post Similar Tweets with Malicious URLs



Post Tweets with Mentions (@ and #)

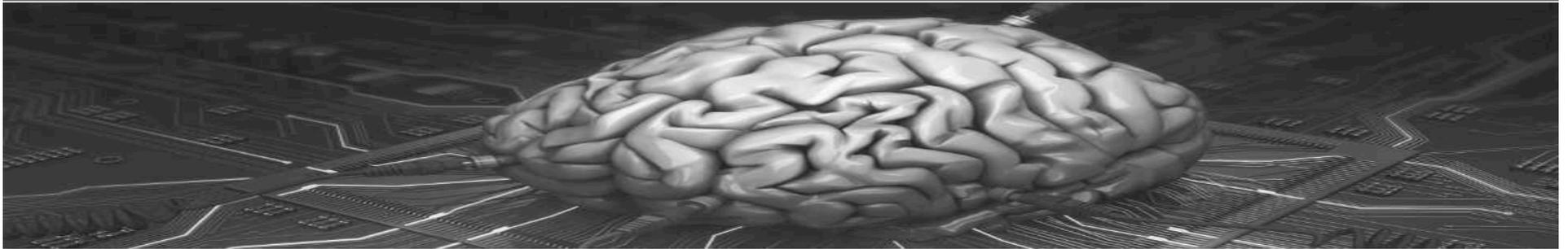


# Roadmap Today

---

- Background
  - **Detecting Malicious OSN Identities**
  - Analyzing Malicious Social Networks
  - Conclusion
-

# Existing Work – Machine Learning Techniques



Label normal and spam accounts  
Design and extract detection features

Profile-based Feature	Content-based Feature
# of Followers	# of Duplicate Tweets
Following to Follower Ratio	Tweet Similarity
# of Tweets	URL Ratio
Reputation	Mention Ratio



# Our Goal

Discover Evasion Tactics

Design New and Robust Detection Features

Formalize Feature Robustness

**u kant c mee**

## Data Collection -- Target

- 
**Twitter spam account:** *“Publish or link to malicious content intended to damage or disrupt other users’ browsers or computers, or to compromise other users’ privacy” -- The Twitter Rules*
- 
*We target this type of spam accounts posting malicious URLs, since these accounts are very parlous and prevalent on Twitter.*

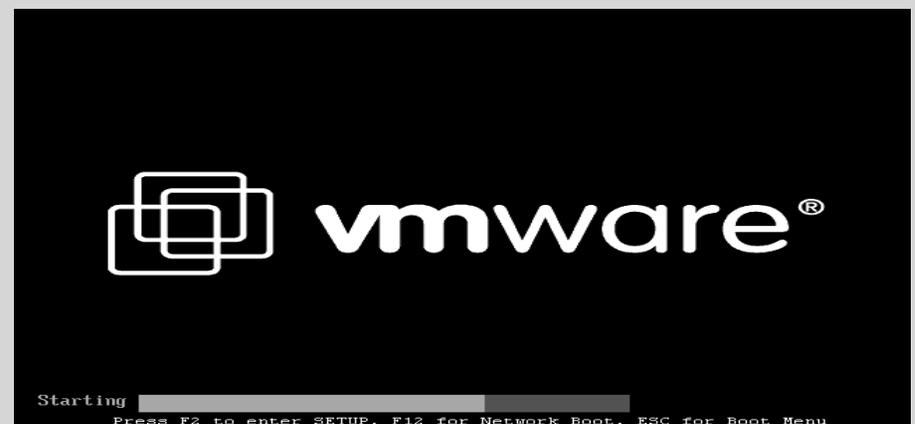
# Data Collection

Item	Value
# of Accounts	485,721
# of Followings	791,648,649
# of Followers	855,772,191
# of Tweets	14,401,157
# of URLs	5,805,351
# of Affected Accounts	10,004
# of Candidate Spam Accounts	2,933
# of Identified Spam Accounts	2,060

## Blacklist Detector



## Honeytrap Detector

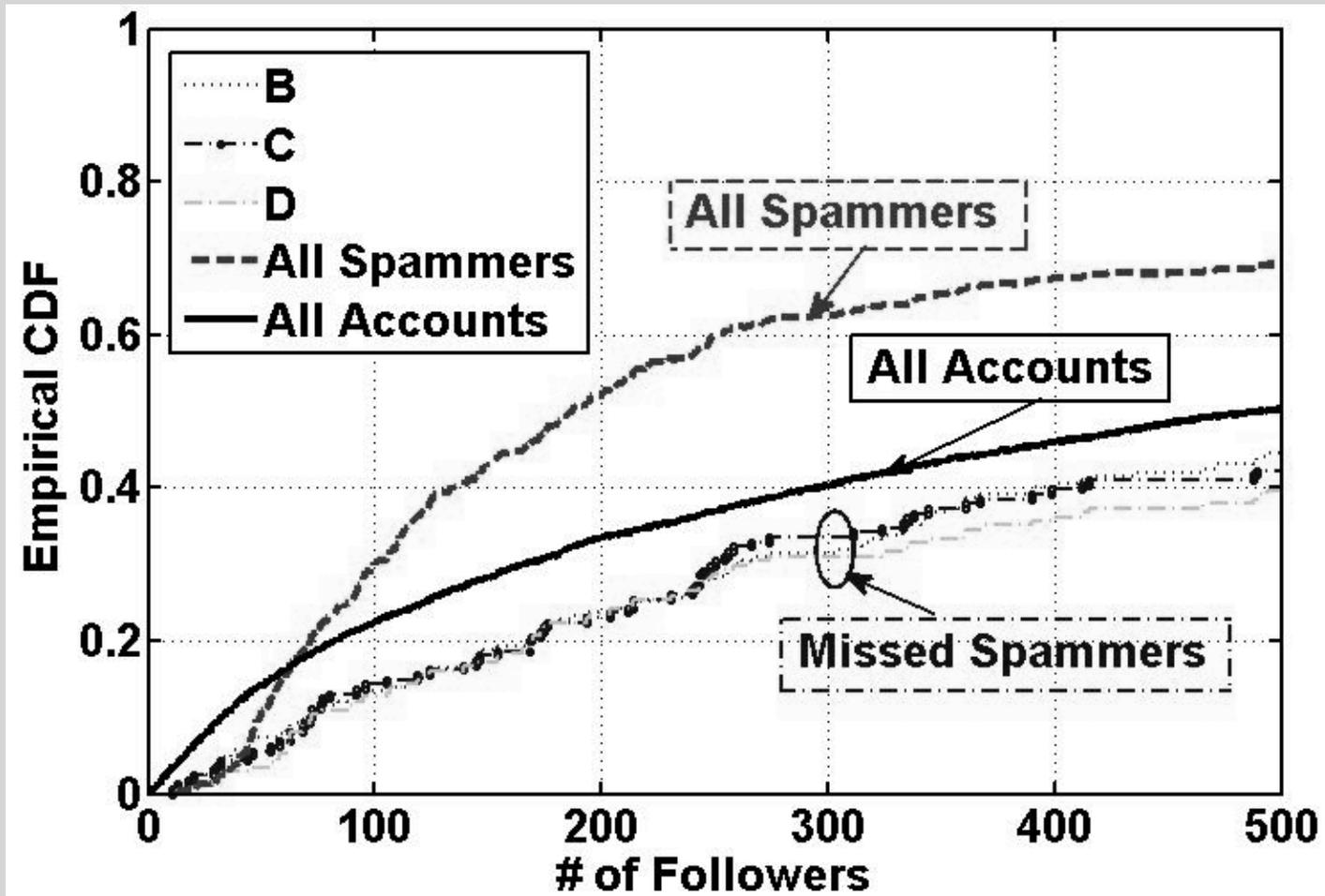


## Examine Existing Work

- ④ Examine three existing work
  - ④ B – Lee et al. [SIGIR' 10]
  - ④ C – Stringhini et al. [ACNSAC' 10 ]
  - ④ D – Wang et al. [SECRYPT' 10 ]
- ④ Extract and analyze spam accounts misclassified as normal accounts (false negatives) in three existing work

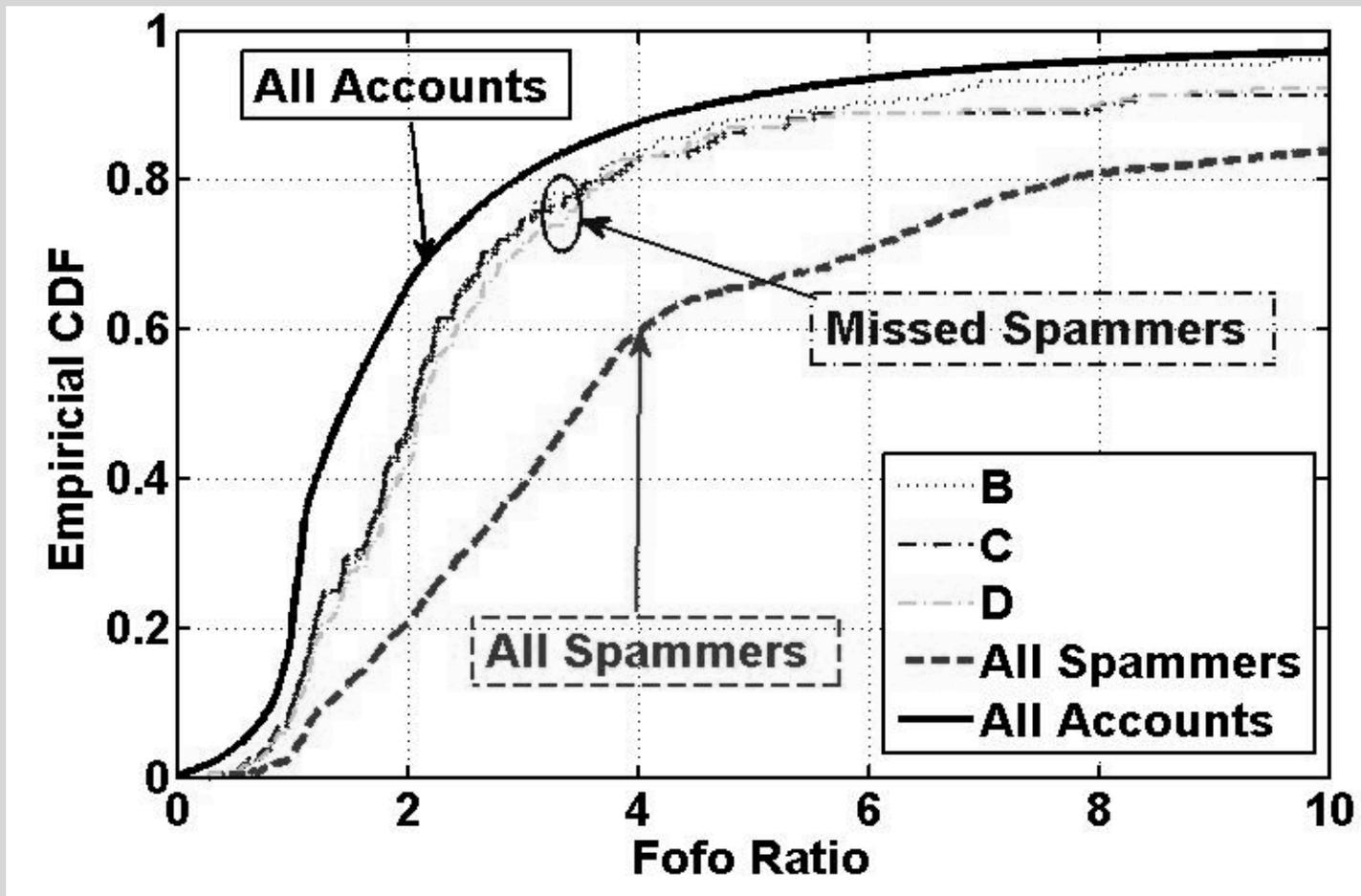
# Analyze Missed Spam Accounts on Existing Work

## # of Followers



# Analyze Missed Spam Accounts on Existing Work

## Following to Follower Ratio



# Evasion Tactics: Profile-based Feature Evasion Tactics

Gaining More Followers

# of Followers

Following to Follower Ratio



<p>Twitter 1000 Plus Package <b>\$24.97</b> Followers delivered in 10-14 days <b>Buy Now</b></p>	<p>Twitter 2000 Plus Package <b>\$44.97</b> Followers delivered in 20-30 days <b>Buy Now</b></p>	<p>Twitter 5000 Plus Package <b>\$79.97</b> Followers delivered in 45-60 days <b>Buy Now</b></p>
<p>1000 Targeted Followers <b>\$34.97</b> Followers delivered in 10-14 days <b>Buy Now</b></p>	<p>2000 Targeted Followers <b>\$54.97</b> Followers delivered in 20-30 days <b>Buy Now</b></p>	<p>5000 Targeted Followers <b>\$89.97</b> Followers delivered in 45-60 days <b>Buy Now</b></p>

Posting More Tweets

# of Tweets

# Evasion Tactics: Content-based Feature Evasion Tactics

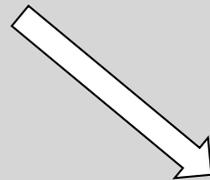
Mixing Normal Tweets



# of Tweets

URL Ratio

Tweet Similarity



The screenshot shows three tweets. The first tweet is about signal proteins from pearls. The second tweet is a promotional offer for a free Twitter account, with 'FREE Twitter' and the URL 'http://bit.ly/6z9ksY' circled. The third tweet is a quote by Thomas Jefferson, with the entire text 'Health is worth more than learning.--Thomas Jefferson' circled.

Signal proteins from pearls can stimulate new skin and bone regeneration <http://bit.ly/2eX5kB>  
30 Dec 09

Get your **FREE Twitter** account and watch your tweets go viral! <http://bit.ly/6z9ksY> #tweetreturns  
11 Dec 09

Health is worth more than learning.--Thomas Jefferson  
21 Aug 09

# Evasion Tactics: Content-based Feature Evasion Tactics

Posting Heterogeneous Tweets

Tweet Similarity



Check out <http://tinyurl.com> . I will get more . You can too!  
27 Apr 10

Every person checks out <http://quu.nu> , you will get more .  
26 Apr 10

want get more, you need to check <http://quu.nu> .  
25 Apr 10

## SpinBot

Free, text spinning, word rewriting, automatic creativity engine.  
Enter English Text to Spin:

spammers are not nice people

Process Text Spin Capitalized Words  Ignore Words Containing:

Text With a Spin:

spammers are not fantastic folks

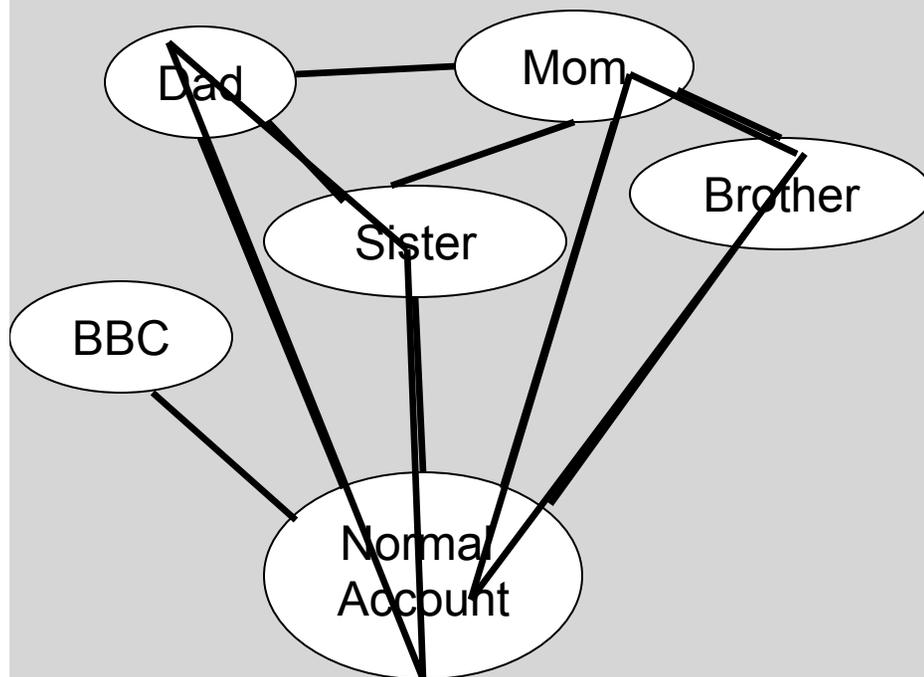
# Designing New and Robust Features

-  Graph-Based Features
-  Neighbor-based Features
-  Automation-based Features
-  Timing-based Features

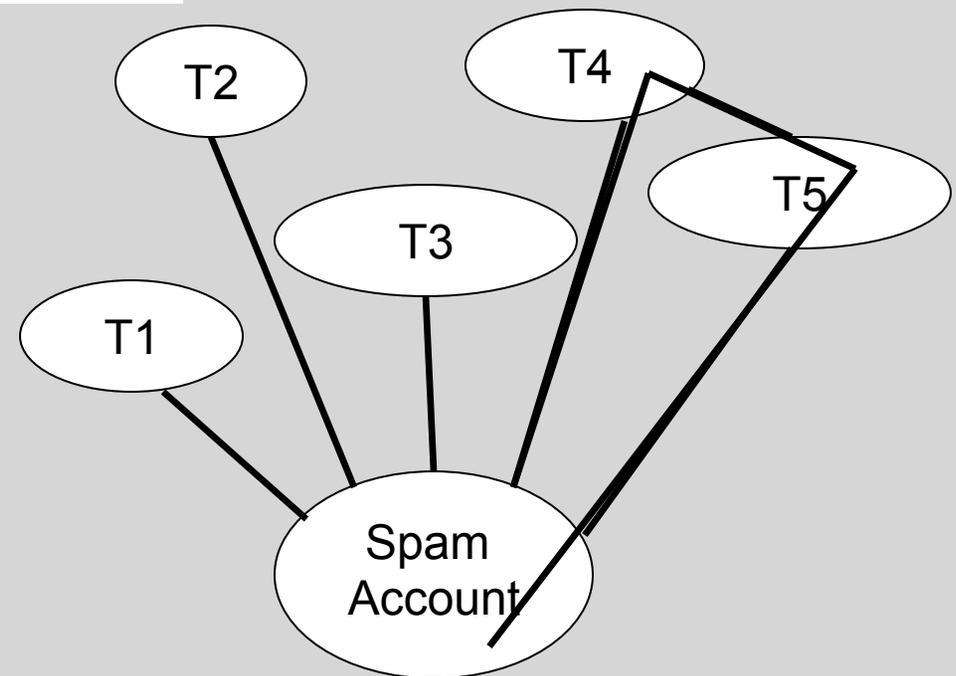
# Graph-Based Features

## Local Clustering Coefficient:

$$LC(v) = \frac{2|e^v|}{K_v \cdot (K_v - 1)}$$



Many Triangles

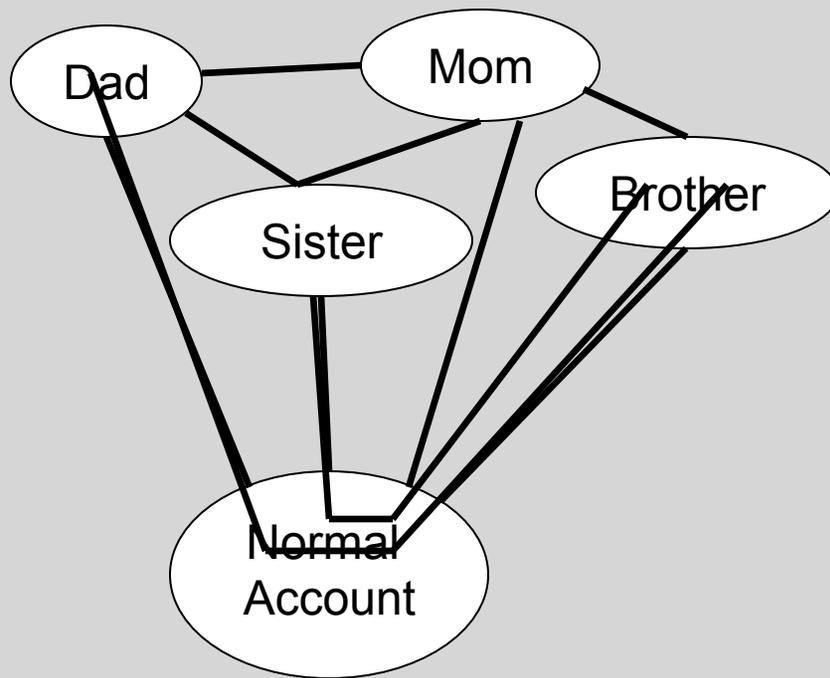


Few Triangles

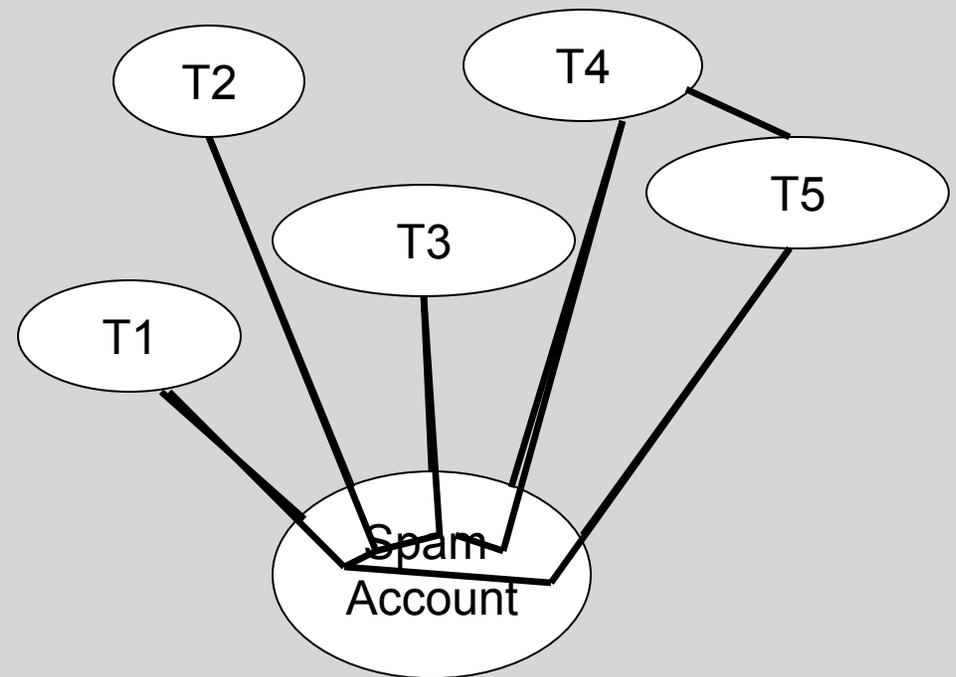
# Graph-Based Features

## Betweenness Centrality:

$$BC(v) = \frac{1}{(n-1)(n-2)} \cdot \sum_{s \neq v \neq t \in V} \frac{\delta_{st}(v)}{\delta_{st}}$$



Few shortest paths passing

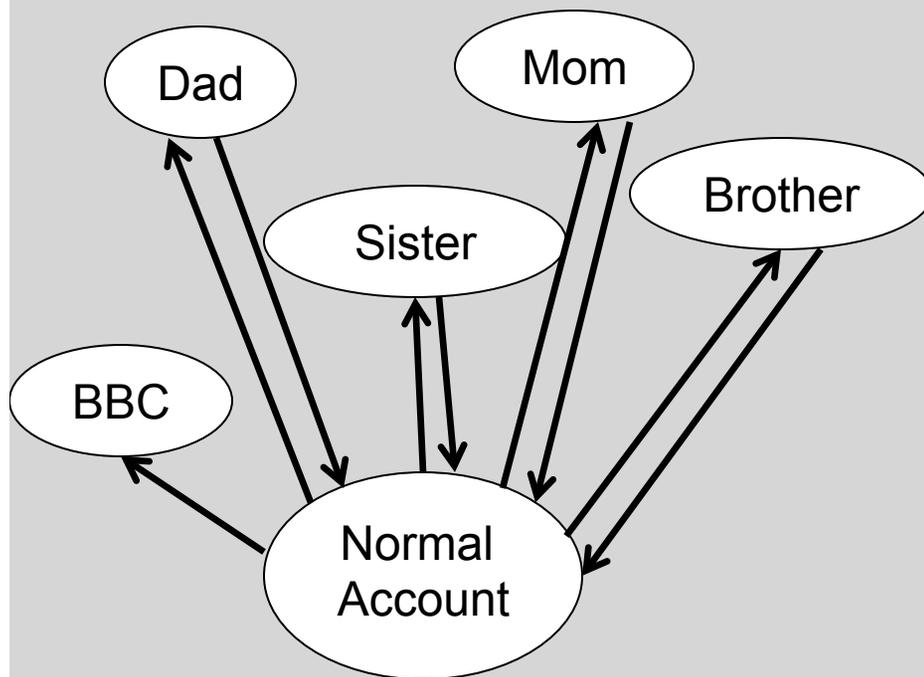


Many shortest paths passing

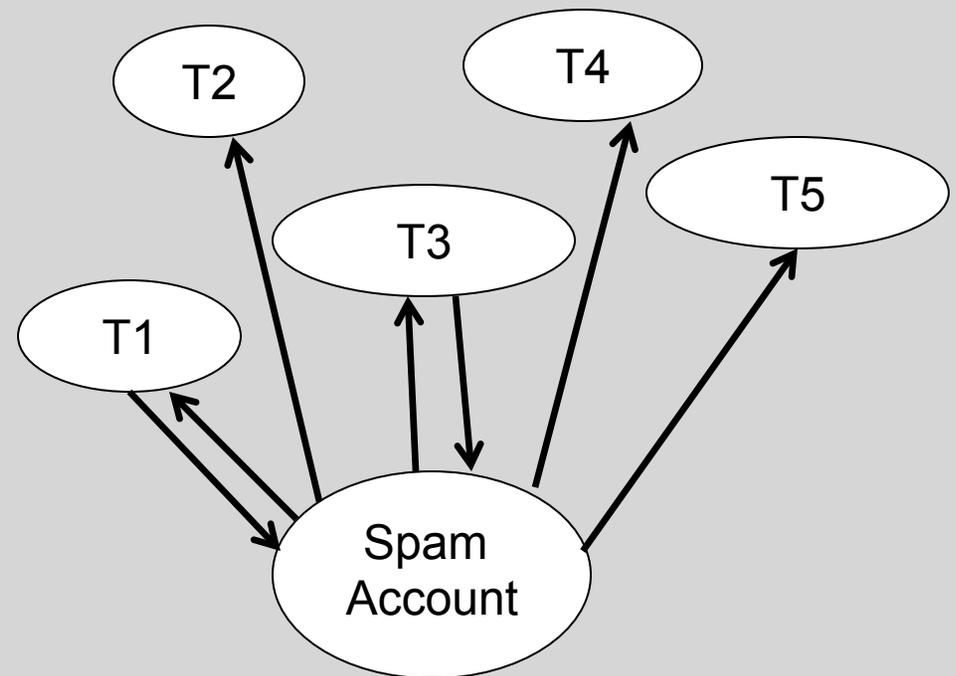
# Graph-Based Features

## Bi-directional Links Ratio:

$$R_{bilink} = \frac{N_{bilink}}{N_{fing}}$$



Ratio = 4/5 = 80%

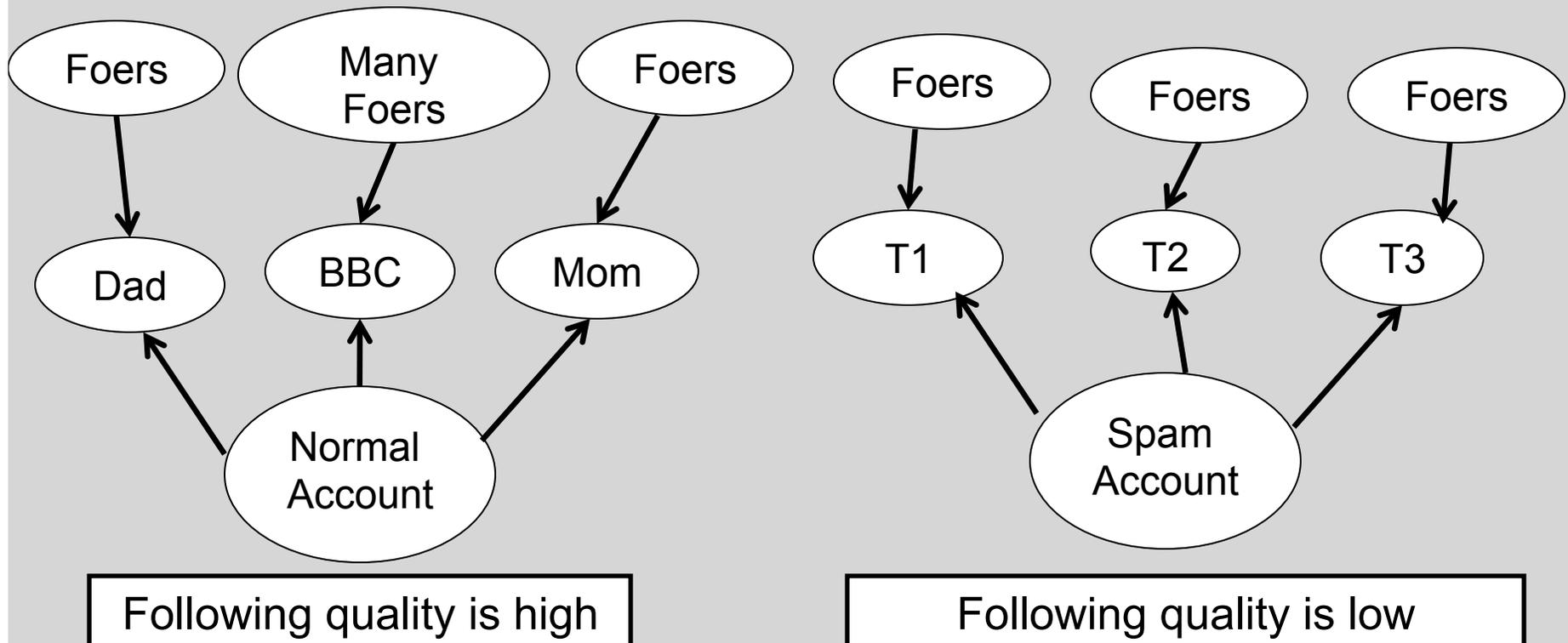


Ratio = 2/5 = 40%

# Neighbor-Based Features

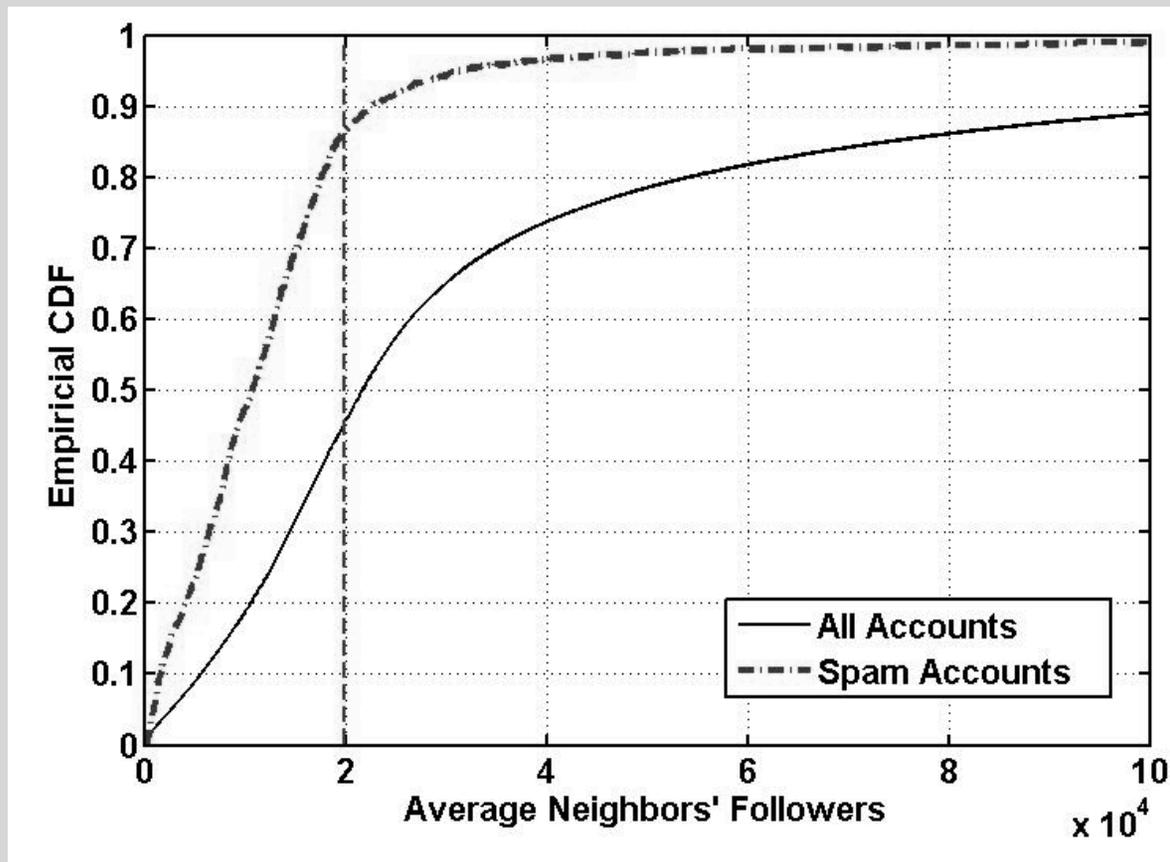
Average Neighbors' Followers:

$$A_{nfer}(v) = \frac{1}{|N_{fing}(v)|} \cdot \sum_{u \in N_{fing}(v)} N_{fer}(u)$$



# Neighbor-Based Features

## ④ Average Neighbors' Followers



## ④ Average Neighbors' Tweets

# Automation-based Features

## Intuition

- Many spammers utilize customized and automated spamming tools designed using Twitter API to post malicious tweets. Especially, if a spammer maintains multiple spam accounts, it will be expensive to organize them to post malicious tweets only manually.

## Features

- API Ratio
- API URL Ratio
- API Similarity



# Formalizing Feature Robustness

- Formalizing the Robustness
  - In order to be robust, a feature must be either expensive or difficult to evade
  - Tradeoff between the spammers' cost  $C(F)$  to evade the detection and the profits  $P(F)$

$$R(F) = C(F) - P(F)$$

- Note: please refer to our RAID'11 paper for details.*

# Robustness of Profile-based Features

## Robustness of “Following to follower ratio” (F3)

Small

$$R(F_3) = \frac{N_{fing}}{T_{F_3}} \cdot C_{fer} - N_{foing} \cdot P_{fing} \quad (T_{F_3} \geq 1)$$

Website	\$ / Follower	Website	\$ / Follower
BuyTwitterFriends.com	0.0049	SocialKik.com	0.0150
TweetSourcer.com	0.0060	USocial.net	0.0440
UnlimitedTwitterFollowers.com	0.0074	Tweetcha.com	0.0470
PurchaseTwitterFollowers.com	0.0490	Twitter1k.com	0.0209

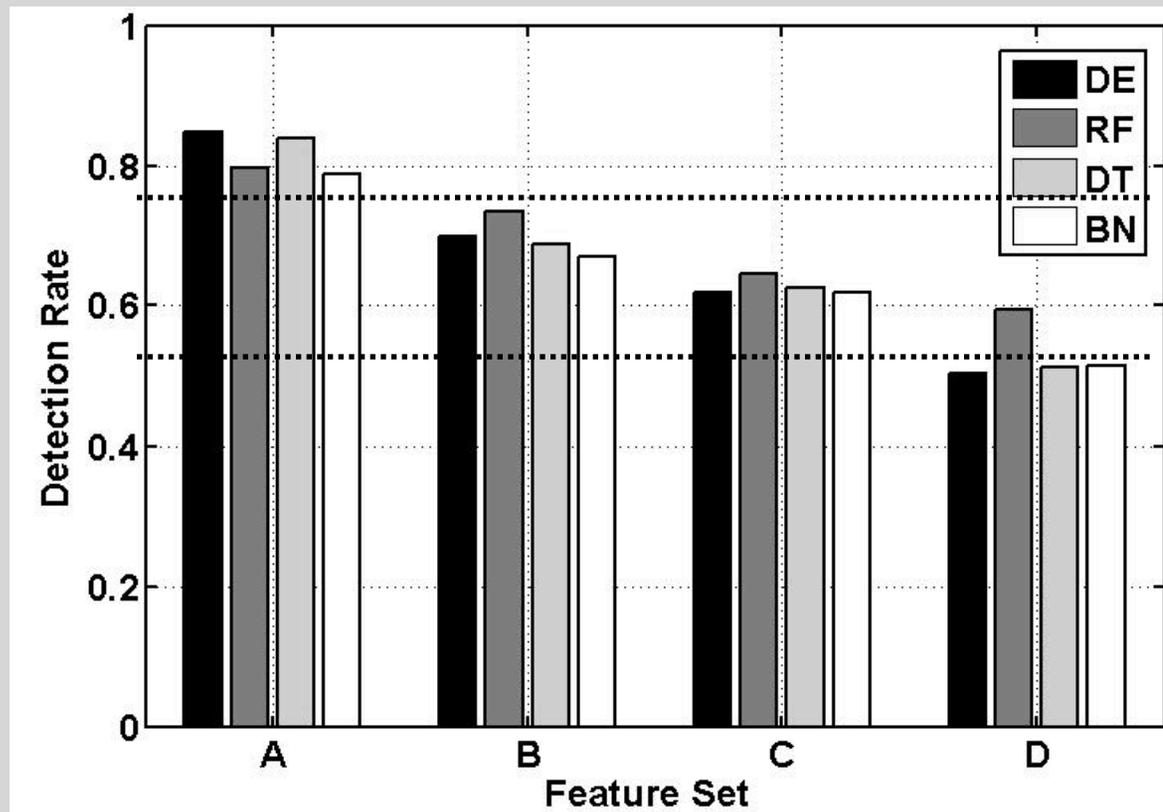
Similar conclusions can be drawn for the features such as “# of followers” and “following to follower ratio”.

# Evaluation

- ④ Feature Set: 8 existing effective features and 10 newly designed features
- ④ Machine Learning Classifier:
  - ④ *Decorate (DE)* , *Random Forest (RF)*
  - ④ *Decision Tree (DT)* , *Bayes Net (BN)*
- ④ Comparison Work
  - ④ A – Our work; B – Lee et al. [SIGIR' 10]
  - ④ C – Stringhini et al. [ACSAC' 10]; D – Wang et al. [SECRYPT' 10]
- ④ Two Data set
  - ④ Data Set I: 5,000 normal accounts and 500 spam accounts
  - ④ Data Set II: 3,500 unlabeled accounts

# Performance Comparison

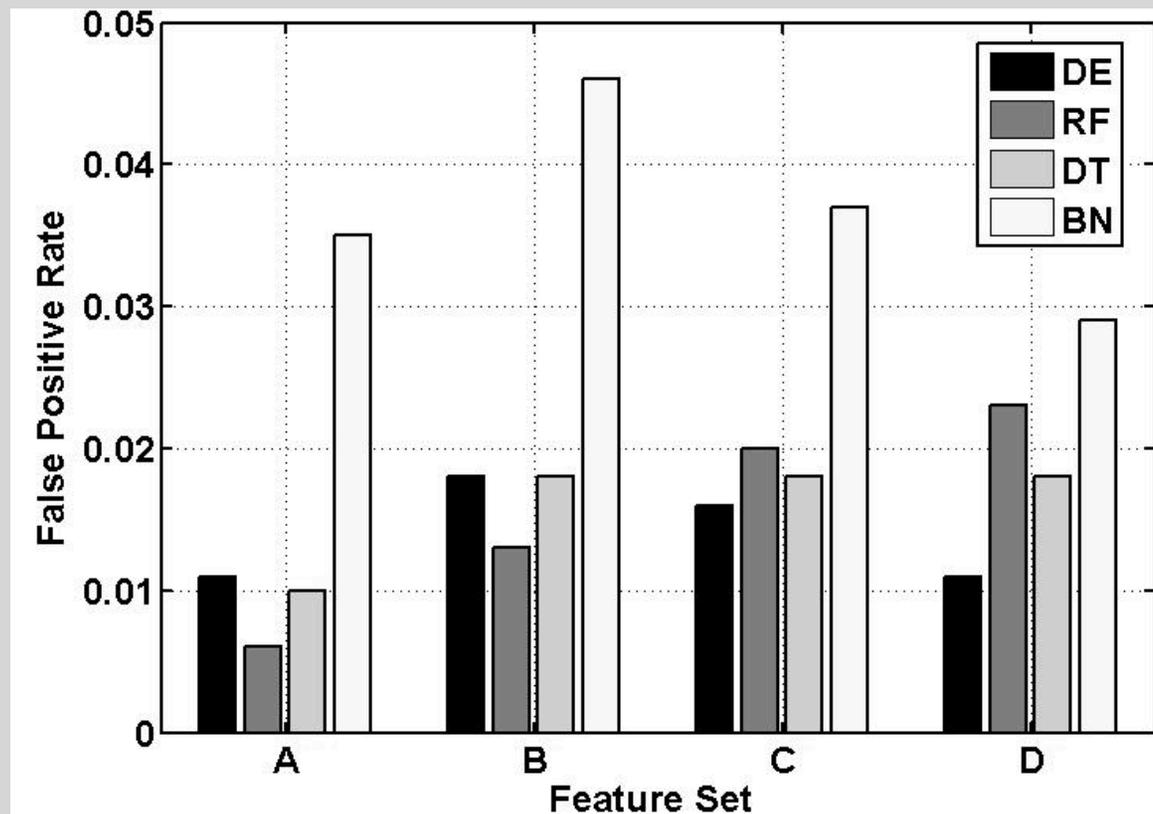
## Detection Rate



- Our best is 86%; In other work, the worst is 51% and the best is 73%.

# Performance Comparison

## False Positive Rate



- Our best performance is 0.5%, which is around half of that of the best performance in three existing work.

# Feature Validation

- Ⓢ Without New Features: 8 existing features
- Ⓢ With New Features: 8 existing + 10 new features
- Ⓢ Detection Rate (DR), False Positive Rate (FPR), F-Measure (FM)

Algorithm	Without New Features			With New Features		
	DR	FPR	FM	DR	FPR	FM
DE	73.8%	1.7%	0.774	85.8%	1.0%	0.877
RF	72.8%	1.2%	0.786	83.6%	0.6%	0.884
DT	70.2%	1.5%	0.757	84.6%	1.1%	0.866
BN	64.4%	4.0%	0.730	78.4%	2.3%	0.777

## Evaluation: Data Set II

- ④ Newly crawl 3,500 unlabeled accounts
- ④ Used the detector trained on the first data set and use Decorate to classify
- ④ Bayesian detection rate of 88.6% (62/70), 17 accounts post malicious URLs detected by Google Safe Browsing blacklist

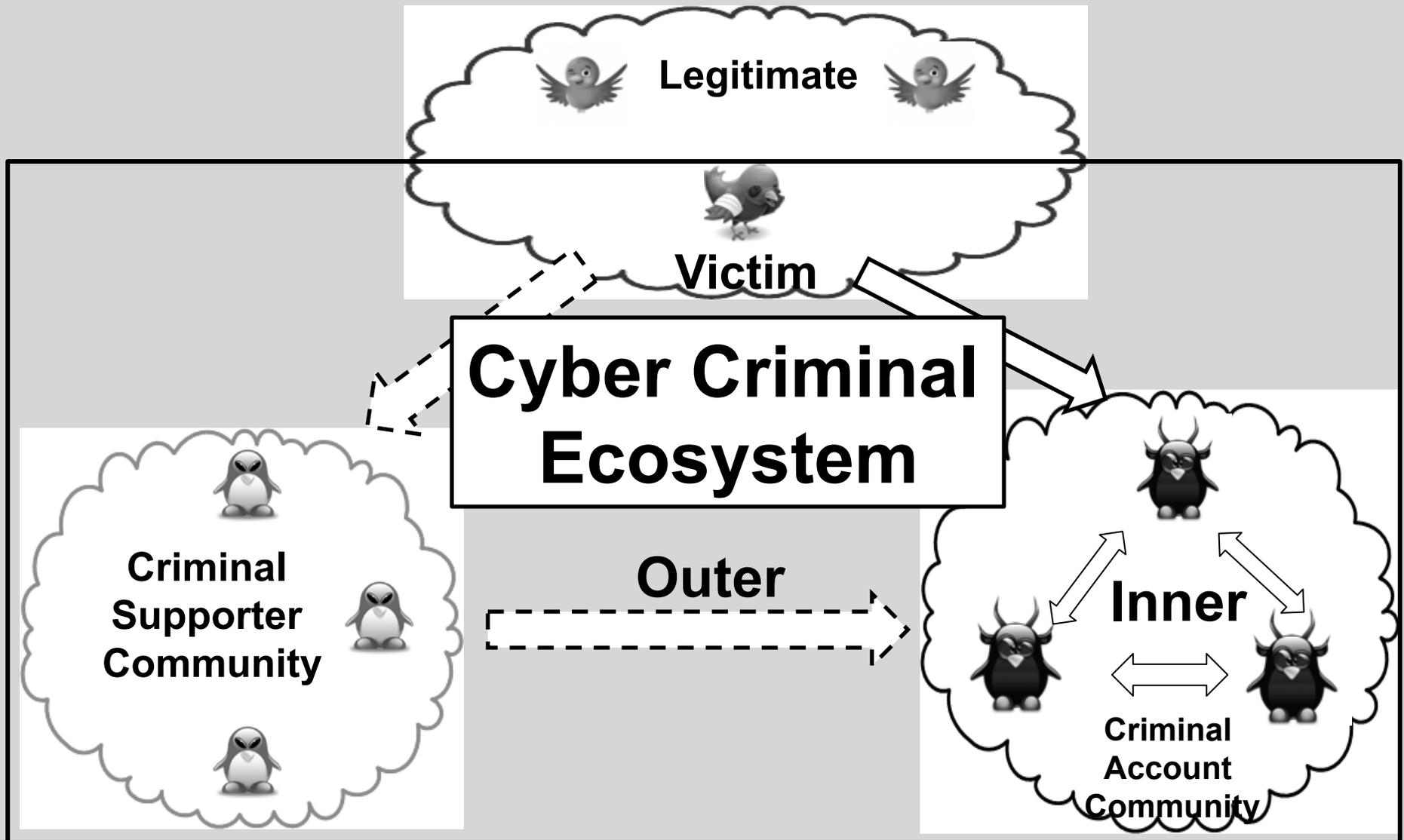
Item	Value
Total Spammer Predictions	70
Verified Spammers	37
Promotional Advertisers	25
Benign	8

# Roadmap Today

---

- Background
  - Detecting Malicious OSN Identities
  - **Analyzing Malicious Social Networks**
  - Conclusion
-

# Background: Cyber Criminal Ecosystem





## Research Goals

**We try to Answer**

**WHAT IS THE STRUCTURE OF CRIMINAL ACCOUNTS' NETWORK?**

**WHAT ARE POSSIBLE FACTORS AND REASONS LEADING TO THAT STRUCTURE?**

**WHAT ARE TYPICAL CHARACTERISTICS OF CRIMINAL SUPPORTERS?**

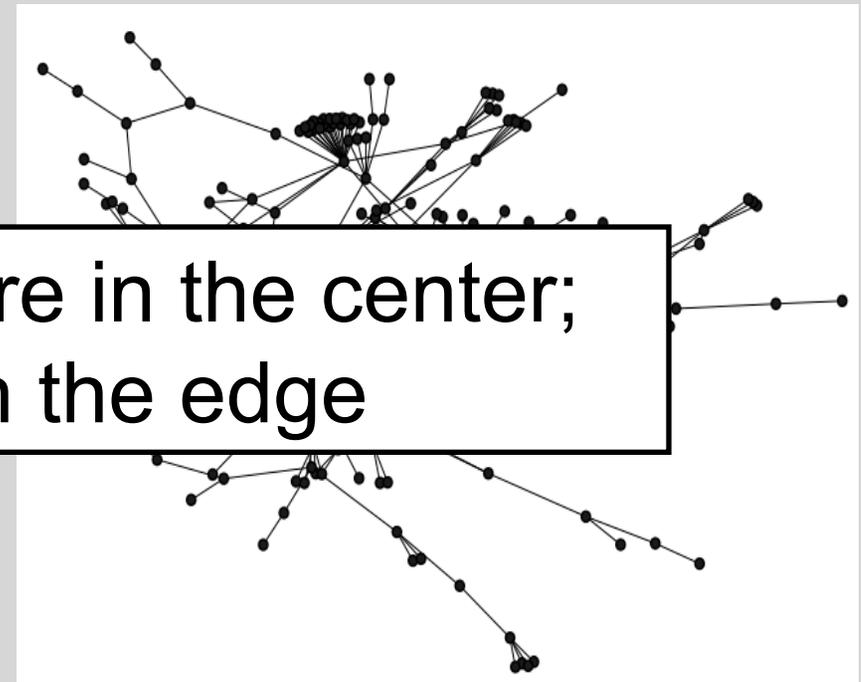
**CAN WE DESIGN NEW DEFENSE ALGORITHMS TO CATCH MORE CRIMINAL ACCOUNTS?**

**AND SO ON ...**

# Inner Relationships: Visualizing Relationship Graph

Node: each criminal account

Edge: Criminal accounts tend to be socially connected



Some accounts are in the center;  
some are in the edge

# Inner Relationships: Revealing Relationship Characteristics

- Observation 1: Criminal accounts tend to be socially connected, forming a small-world network
  - Graph Density:  $\frac{|E|}{|V| \cdot (|V| - 1)}$ 
    - Criminal graph:  $2.33 \times 10^{-3}$
    - Public Twitter snapshot (41.7m nodes and 1.47b edges):  $8.45 \times 10^{-7}$
  - Average Shortest Path Length
    - Criminal graph: 2.60
    - Public Twitter snapshot (3,000 nodes): 4.12
  - Reciprocity: 95% criminals are higher than 0.2; 55% normal accounts are higher than 0.2

## Explanations

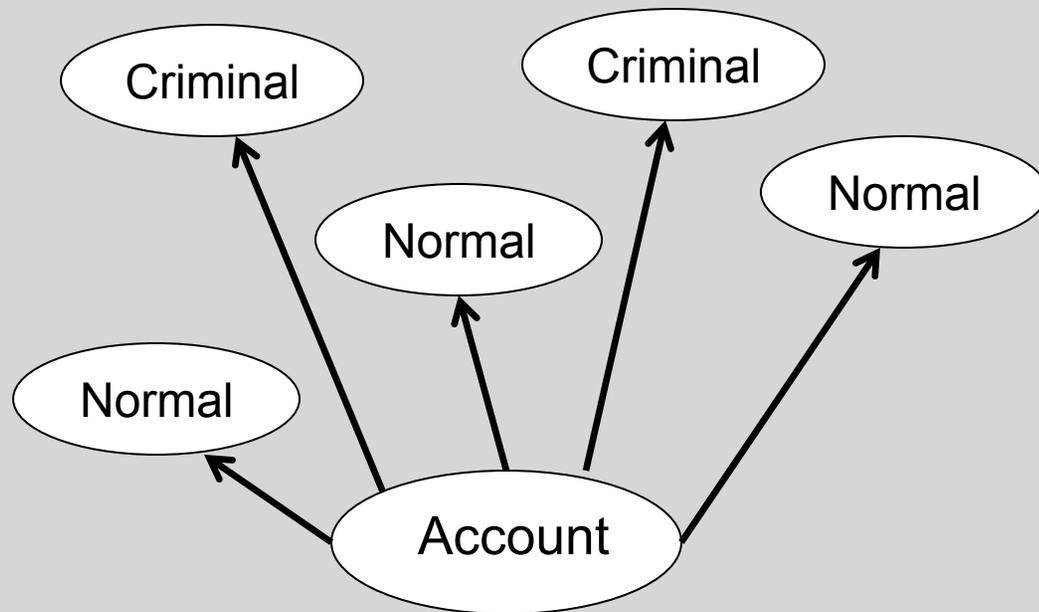
- ④ Criminal accounts tend to follow many other accounts without considering those accounts' quality much, making themselves to connect to other criminal accounts.
- ④ Criminal accounts, belonging to the same criminal organizations, may be artificially/intentionally connected with each other.

# Inner Relationships: Revealing Relationship Characteristics

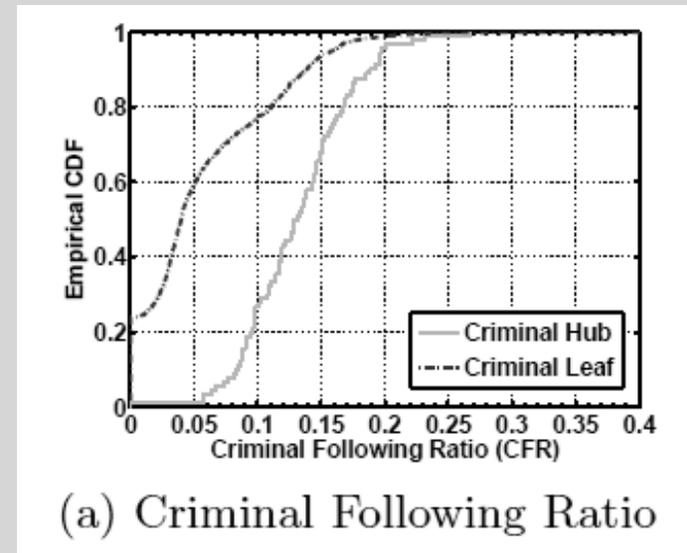
- Observation 2: Compared with criminal leaves (nodes at the edge), criminal hubs (nodes in the center) are more inclined to follow criminal accounts.
  - Extract hubs and leaves: HITS algorithm
  - K-means: 90 hubs, 1970 leaves

# Cont.

- Calculate Criminal Following Ratio (in our collected Twitter snapshot)



Ratio =  $2/5 = 40\%$



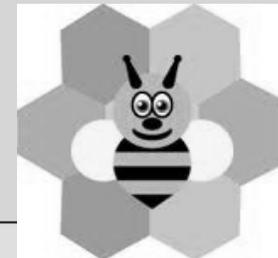
**Hubs tend to follow criminal accounts**

## Cont.

- Similar to the Bee Community, in the criminal account community, criminal leaves, like worker bees, mainly focus on collecting pollen (randomly following other accounts to expect them to follow back)



- Criminal hubs in the interior, like queen bees, mainly focus on supporting bee workers and acquiring pollen from them (following leaves and acquiring their followers' information).



## Outer Social Relationships

-  If criminal accounts mainly build inner social relationships within themselves, criminal accounts can be easily detected.
-  However, Twitter criminal accounts have already utilize several tricks to obtain followers outside the criminal account community and mix well into the whole Twitter space.
-  Criminal Supporters
  -  outside the criminal community
  -  have close “follow relationships” with criminal accounts

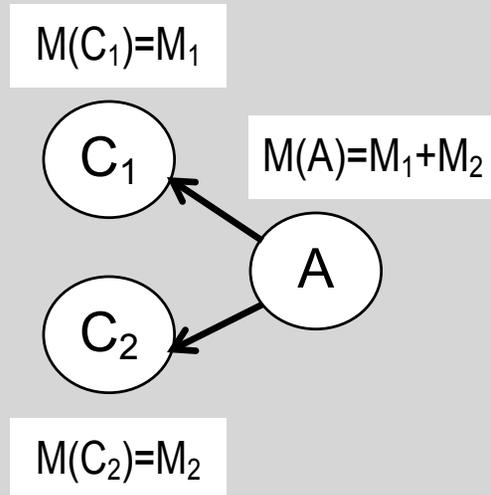
# Outer Social Relationships: Extracting Criminal Supporters

- ④ Malicious Relevance Score Propagation Algorithm (Mr.SPA)
  - ④ Assign a malicious relevance score to measure social closeness to criminals
  - ④ The more criminal accounts that an account has followed, the higher score should inherit;
  - ④ The further an account is away from a criminal account, the lower score should inherit;
  - ④ The closer the support relationship between a Twitter account and a criminal account is, the higher score should inherit.

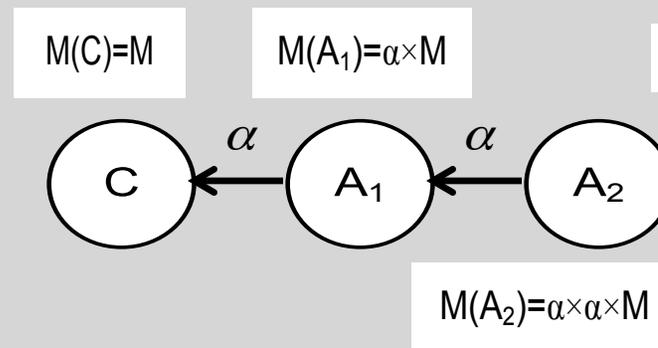
# Outer Social Relationships: Extracting Criminal Supporters

- Score Initialization: assigned a non-zero score to each criminal account
- Score Propagation: based on three intuitions

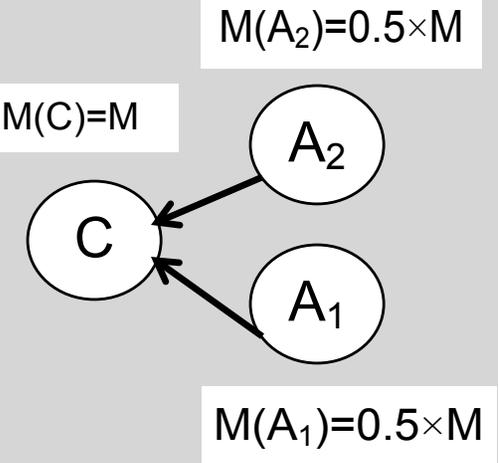
Aggregation



Dampening



Splitting



- Threshold: x-means; 5,924 criminal

# Outer Social Relationships: Characterizing Criminal Supporters

- ④ **Social Butterflies:** have extraordinarily large numbers of followers and followings
- ④ 3,818 social butterflies
- ④ **Assumption: butterflies tend to follow back the users that first follow themselves without careful examinations.**



## Cont.

- ④ Experiment: examine follow backs
- ④ Create 30 Twitter accounts without any tweets and default registration information

④ 10

④ 10

④ 10

**Social Butterflies tend to automatically follow back any accounts that follow them!**

- ④ Time span: 48 hours

- ④ Result:

- ④ Butterflies: 47.8%
- ④ Normal: 1.8%
- ④ Criminal: 0.6%

# Outer Social Relationships: Characterizing Criminal Supporters

- 🕒 **Social Promoters:** have large following-follower ratios, larger following numbers and relatively high URL ratios.
- 🕒 The owners of these accounts usually use Twitter to promote themselves or their business.
- 🕒 508 social promoters





# Cont.

Assumption: promoters usually promote themselves or their business by actively

The image shows a Twitter profile for a user named 'Concursos' and a screenshot of the website 'CZIP Concursos'. The Twitter profile includes a bio, a 'Follow' button, and a timeline of tweets. The website features a search bar, a category list, and a list of digital course products.

**Twitter Profile:**

- Profile Name: Concursos
- Profile Picture: Open book
- Stats: 5 Tweets, 1,803 Following, 473 Followers, 23 Listed
- Follow Button: + Follow
- Text follow: Text follow [redacted] to 40404 in the United States
- Timeline:
  - Tweet 1: apostilas para concursos do BB, MTE, correios, receita fere, ibge, inss, banco do nordeste <http://bit.ly/alyASj> (11 Feb 10)
  - Tweet 2: tudo para concursos: <http://bit.ly/alyASj> (10 Feb 10)
  - Tweet 3: tudo para concursos: <http://bit.ly/alyASj> (9 Feb 10)

**Website: CZIP Concursos**

- Search Bar: [Search]
- Categories:
  - ABIN
  - Banco do Brasil
  - Banco do Nordeste - BNB
  - Caixa Econômica - CEF
  - Câmara de Salvador
  - Cobra Tecnologia
  - Correios
  - Detran-ES
  - DPE-RJ
  - EMBAKA
  - EMBRATUR
  - ENEM
  - FIOCRUZ
  - IBGE
- Destaque:
  - Product: Apostila Digital Correios 2011 - Agente dos Correios +SIMULADÃO (R\$ 14,90)
- Melhores Produtos:
  - Product 1: Apostila Digital Correios 2011 - Agente dos Correios (R\$ 9,90)
  - Product 2: Apostila Digital Policia Militar PM-SP - Soldado (R\$ 29,90)
  - Product 3: Apostila Digital Policia Militar PM-GA 2011 (R\$ 24,90)
- Informações:
  - Este site é seguro?
  - Quêntas
  - Como Comprar
  - Noticias Concursos
  - Confirmar Pagamento
  - Fale Conosco
- Entrar:
  - E-Mail: [input]
  - Senha: [input]
  - Entrar
  - Criar Conta
  - Esqueceu a senha?
- Sua Compra

**Domain Name Entropy:**

0 1 2 3 4

All ACCOUNTS

# Outer Social Relationships: Characterizing Criminal Supporters

- ④ **Dummies:** are those Twitter accounts who post few tweets but have many followers
- ④ Strange
  - ④ Few tweets
  - ④ Many followers
  - ④ Close to criminal accounts
- ④ 81 dummies



# Cont.

- Assumption: most of dummies are controlled or utilized by cyber criminals.

The MLM BUSINESS WITH A STRING ATTACHED - IT'S FREE!!! <http://is.gd/...9W>

**Killer Software That Makes Me \$3267.75 Every Day Using Twitter!!**

**FREE "Guide To Instant Online Income"**  
 Let Me Show You How To Get Multiple \$100 Payments Everyday Of The Week

Simply enter your primary email address below to start your Free mini-eCourse Today...

Your Name

E-Mail

>> Send My Free Report <<

accounts begin posting (verified) phishing URLs.

- ④ **How can we exploit the malicious social networks?**
- ④ **Given a small seed set of malicious identities, can we infer more?**

# Inferring Criminal Accounts: Main Idea

- ④ Intuitions:
  - ④ Criminal accounts tend to be socially connected;
  - ④ Criminal accounts usually share similar topics (or keywords or URLs) to attract victims, thus having strong semantic coordination among them.
- ④ Criminal account Inference Algorithm (CIA) propagates malicious scores from a seed set of known criminal accounts to their followers according to the **closeness of social relationships** and the **strength of semantic coordination**. If an account accumulates sufficient malicious score, it is more likely to be a criminal account.

# Inferring Criminal Accounts: Design

- ④ The closeness of social relationships
  - ④ Mr. SPA
- ④ The strength of semantic coordination
  - ④ Semantic Similarity score
  - ④ A higher score between two accounts implies that they have stronger semantic coordination
- ④ Infer criminal accounts in a set of Twitter accounts by starting from a known seed set of criminal accounts
- ④ Assign malicious scores for each account based on those two metrics; infer accounts with high malicious scores as criminal accounts

# Inferring Criminal Accounts: Evaluation

## Dataset:

-  Dataset I refers to the one with around half million accounts
-  Dataset II contains another new crawled 30K accounts by starting from 10 newly identified criminal accounts and using breath-first search (BFS) strategy.

## Metric:

-  the number of correctly inferred criminal accounts and malicious affected accounts (denoted as CA and MA, respectively) in a top (ranked) list.

# Inferring Criminal Accounts: Evaluation

## Different Selection Strategies

Selection Size = 4,000

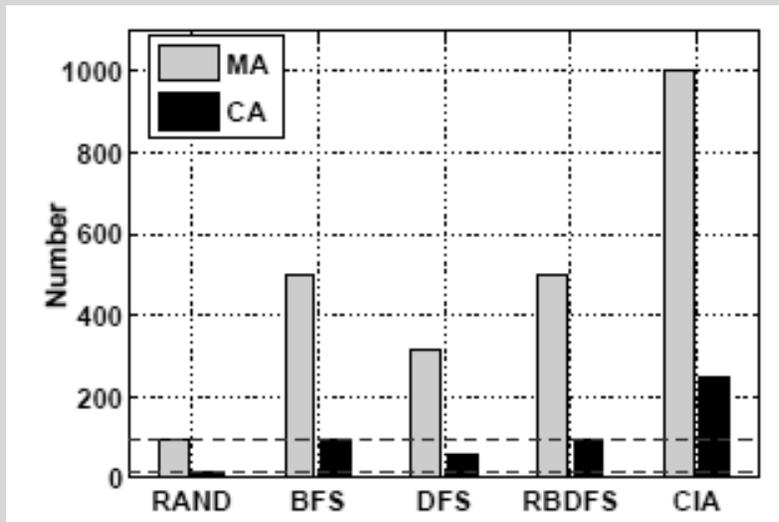
Seed Size = 100

*RAND*: Randomly Select ;

*BFS*: Breath First Search

*DFS*: Depth First Search;

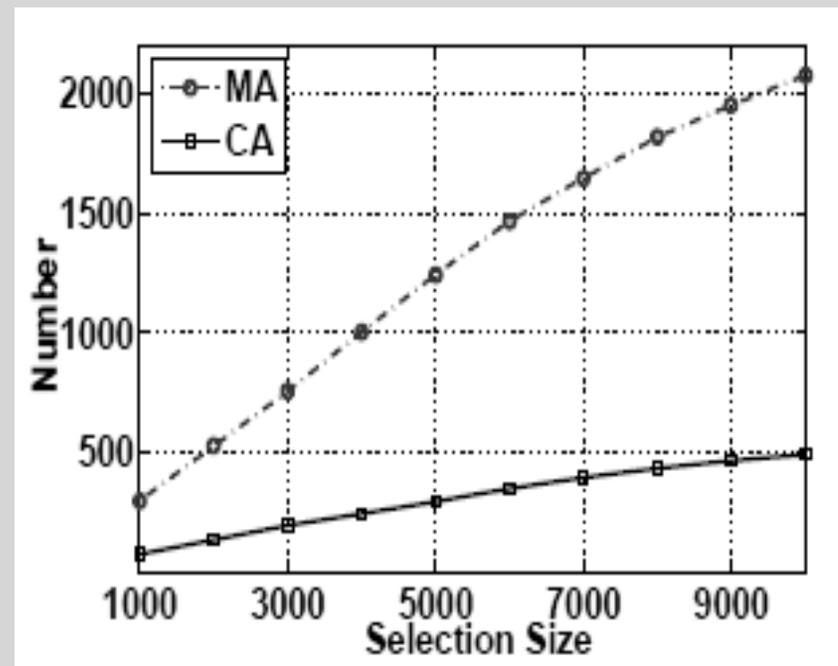
*RBDFS*: Combine *BFS* and *DFS*



## Different Selection Sizes

Selection Strategy: *CIA*

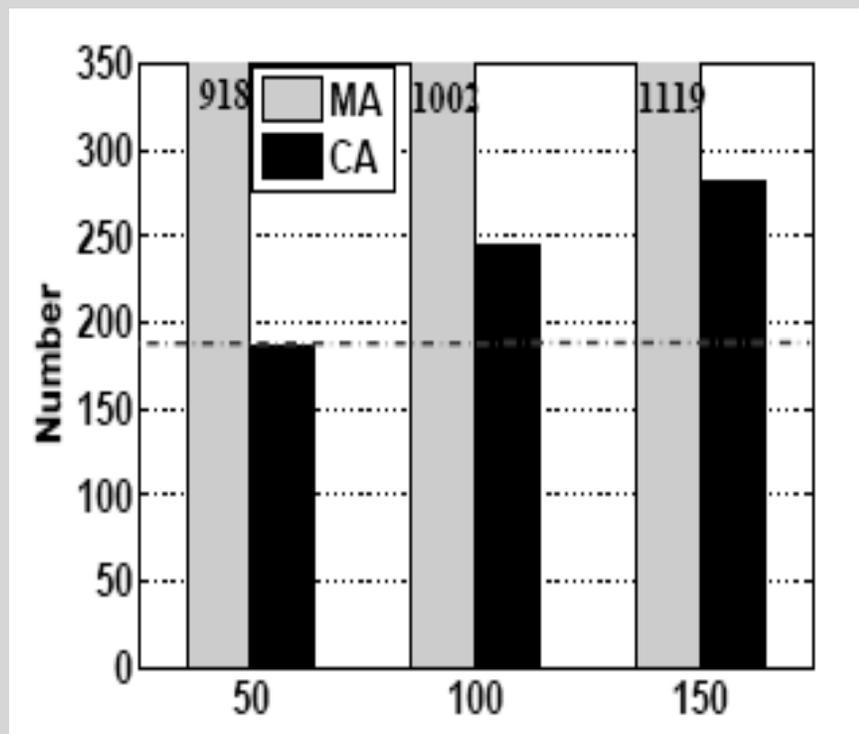
Seed Size = 100



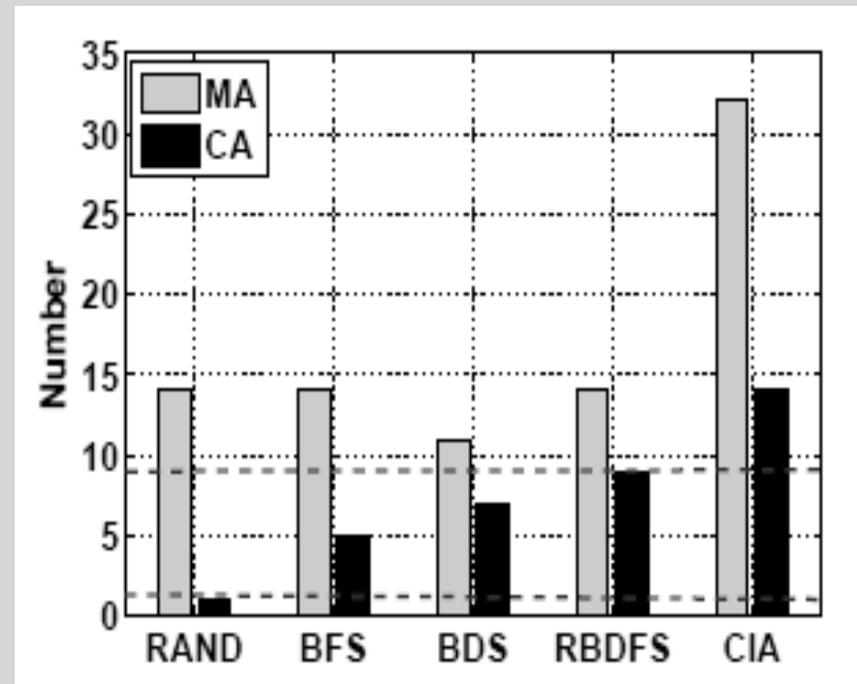
# Cont.

## Different Seed Sizes

Selection Size = 4,000



## Evaluation on Dataset II



More results in Our WWW'12 paper

## Conclusion

- OSN: emerging attack platforms, also a new opportunity to study the community of cyber criminals
- We present
  - New robust features to detect malicious identities
  - The first empirical study of the cyber criminal ecosystem on Twitter
- Can our insights/observations applied to other OSNs?
- Security in social computing/networking is fun...



# Questions & Answers



[Http://faculty.cse.tamu.edu/guofei](http://faculty.cse.tamu.edu/guofei)

## Limitation

- ④ We acknowledge that our analyzed dataset may contain some bias. Also, the number of our analyzed criminal accounts is most likely only a lower bound of the actual number in the dataset, because we only target on one specific type of criminal accounts due to their severity and prevalence on Twitter.
- ④ We also acknowledge that our validations on some possible explanations proposed in this work may be not absolutely rigorous, due to the difficulties in thoroughly obtaining criminal accounts' social actions or motivations.