Machine Learning Meets Social Networking Security:

Detecting and Analyzing Malicious Social Networks for Fun and Profit

Guofei Gu

Se**cu**re **C**ommunication and **C**omput**e**r **S**y**s**tems (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

TEXAS A&M UNIVERSITY

Introduction: OSNs are Popular

TEXAS A&M UNIVERSITY

Background: OSNs are Suffering

TEXAS A&M UNIVERSITY

Backgrounds: Attacks on Twitter

- Detect and suspend malicious OSN accounts individually
- Understand how criminal accounts survive and work on Twitter
- Analyze OSN criminal accounts' social relationships and ecosystem

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

TEXAS A&M UNIVERSITY

Follow Many Accounts

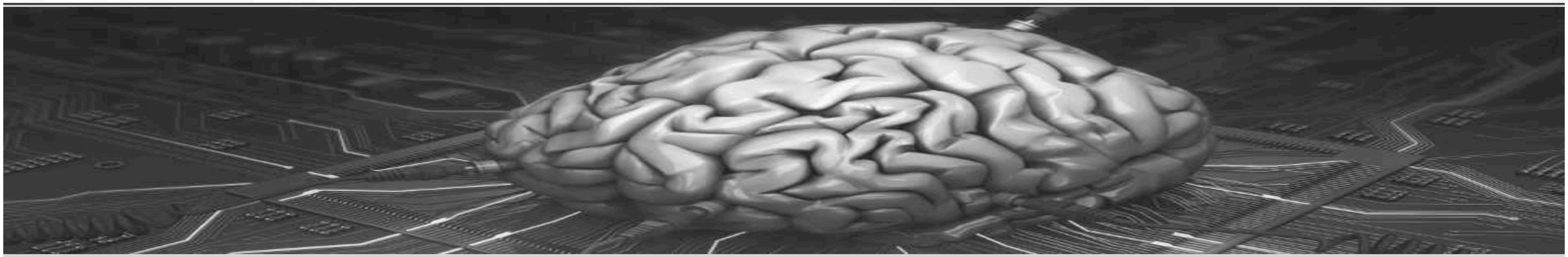
Post Similar Tweets with Malicious URLs

Post Tweets with Mentions (@ and #)

Roadmap Today

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

Existing Work – Machine Learning Techniques

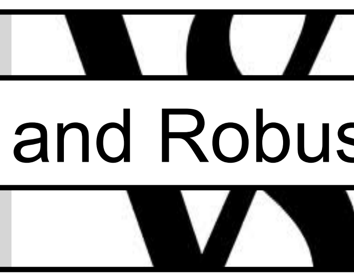


Label normal and spam accounts

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

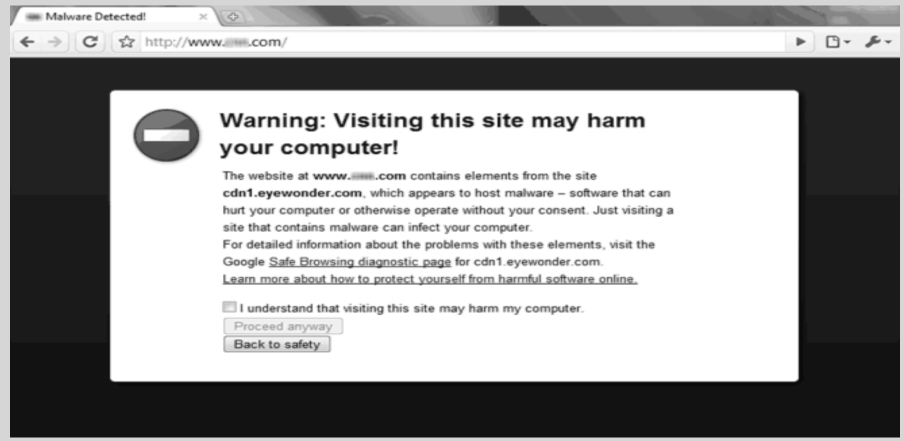
Data Collection -- Target

- **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.*

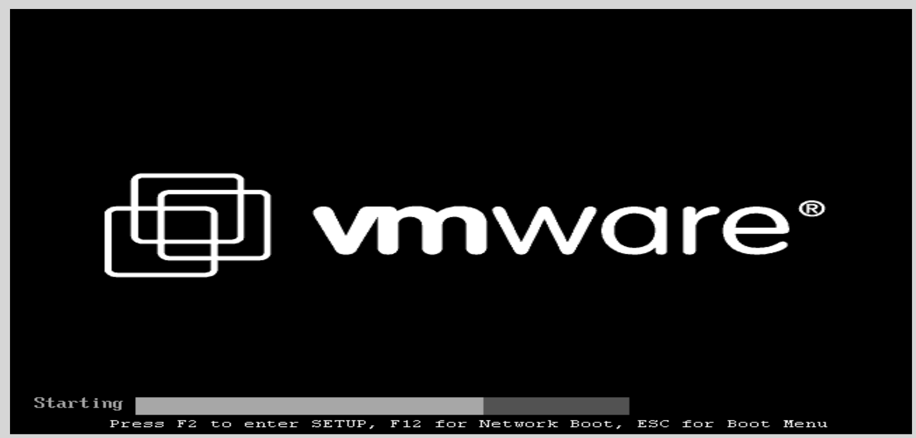
Data Collection

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



Honeypot Detector



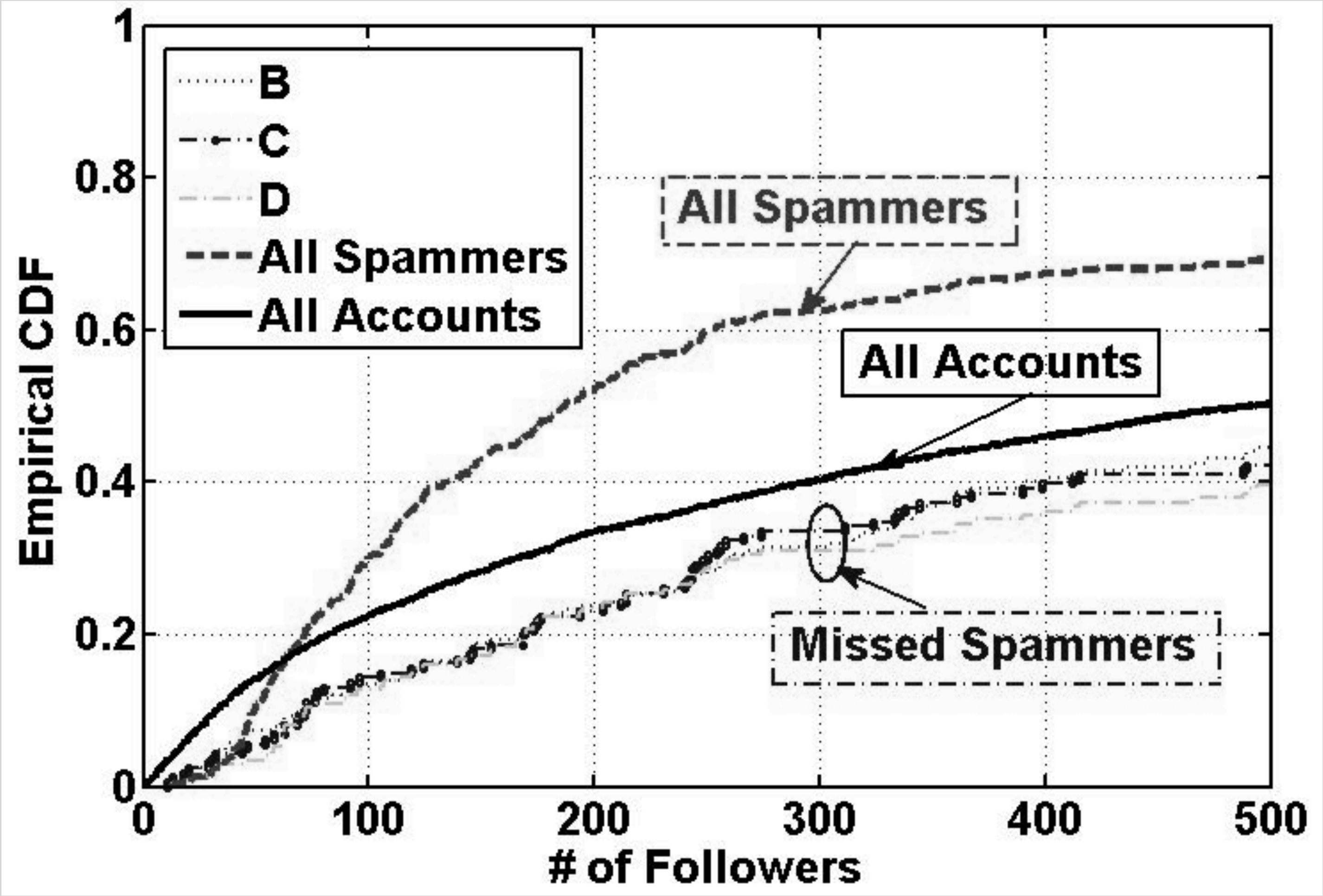
Examine Existing Work

- Examine three existing work
  - B – Lee et al. [SIGIR' 10]
  - C – Stringhini et al. [ACSAC' 10]
  - D – Wang et al. [SECRYPT' 10]

- Extract and analyze spam accounts misclassified as normal accounts (false negatives) in three existing work

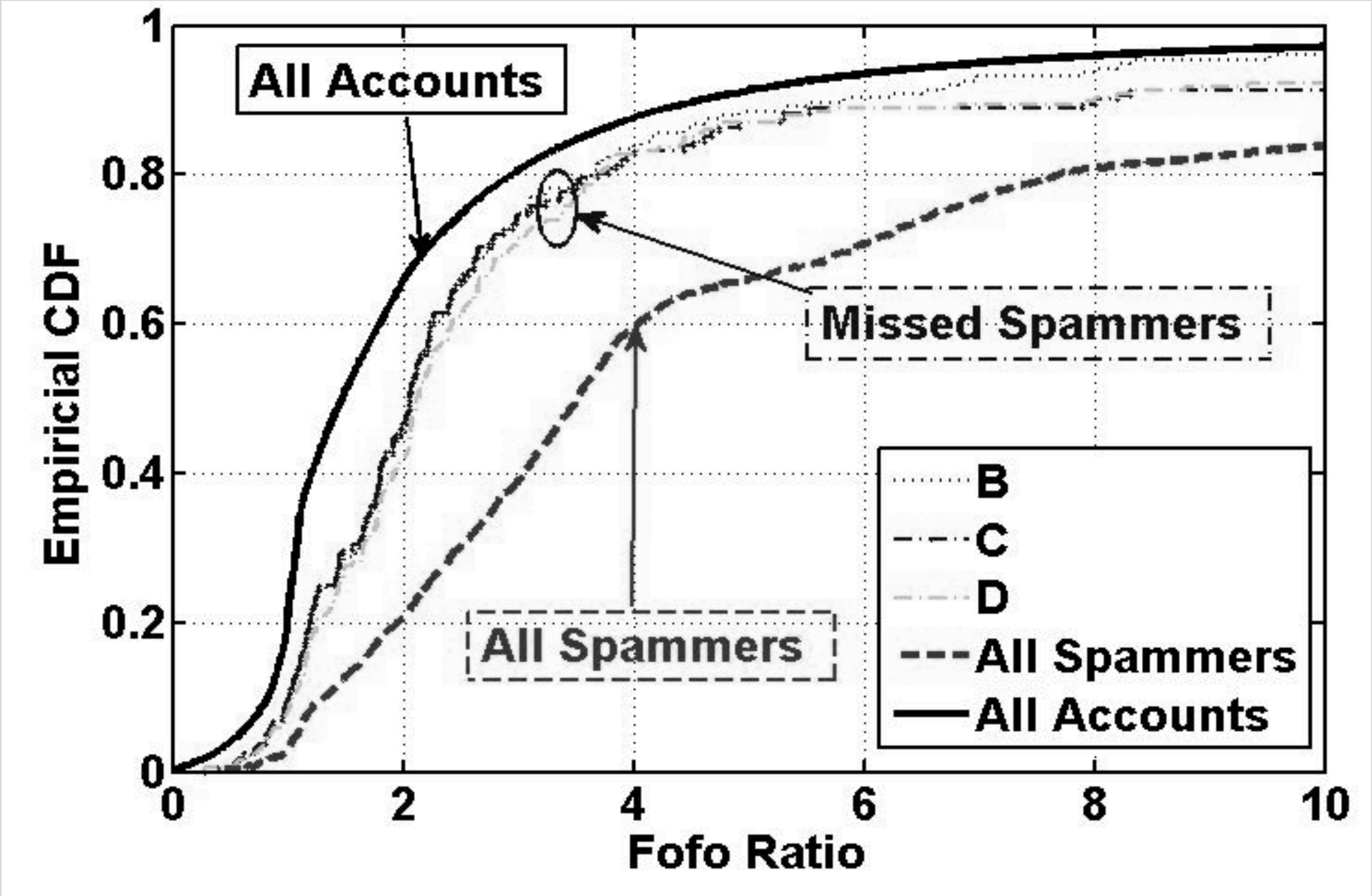
Analyze Missed Spam Accounts on Existing Work

- # of Followers



Analyze Missed Spam Accounts on Existing Work

- Following to Follower Ratio



Evasion Tactics: Profile-based Feature Evasion Tactics

TEXAS A&M UNIVERSITY

Gaining More Followers → # of Followers; Following to Follower Ratio

Posting More Tweets → # of Tweets

Evasion Tactics: Content-based Feature Evasion Tactics

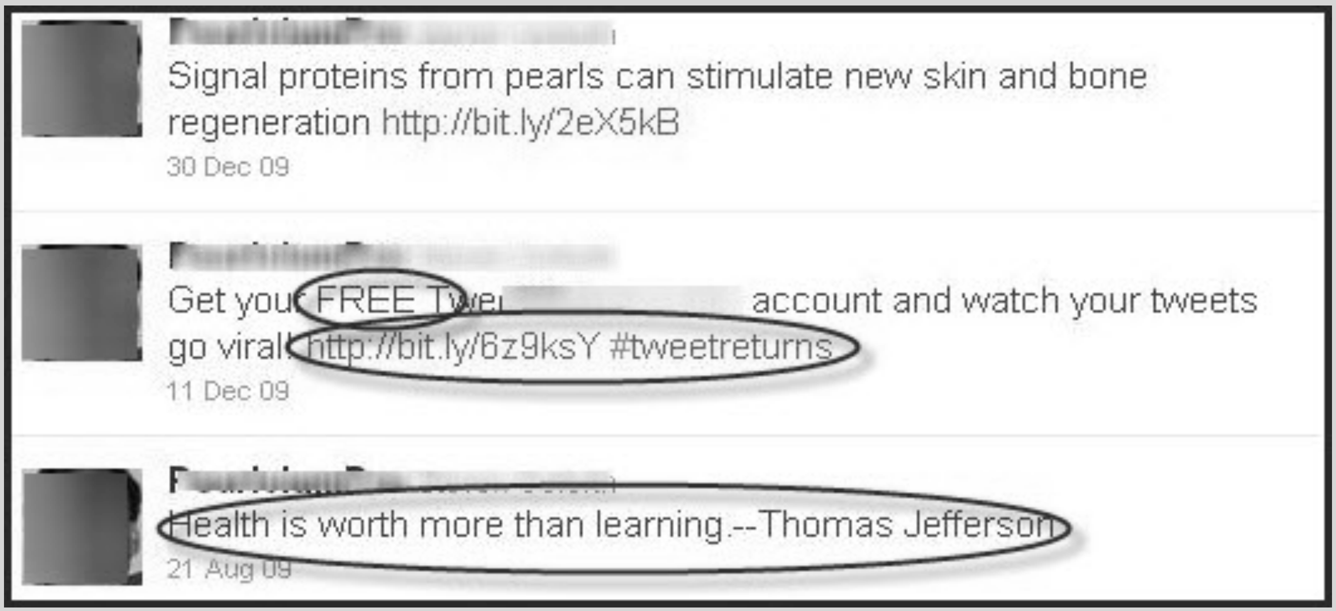
TEXAS A&M UNIVERSITY

Mixing Normal Tweets

# of Tweets

URL Ratio

Tweet Similarity



Evasion Tactics: Content-based Feature Evasion Tactics

Posting Heterogeneous Tweets

Tweet Similarity

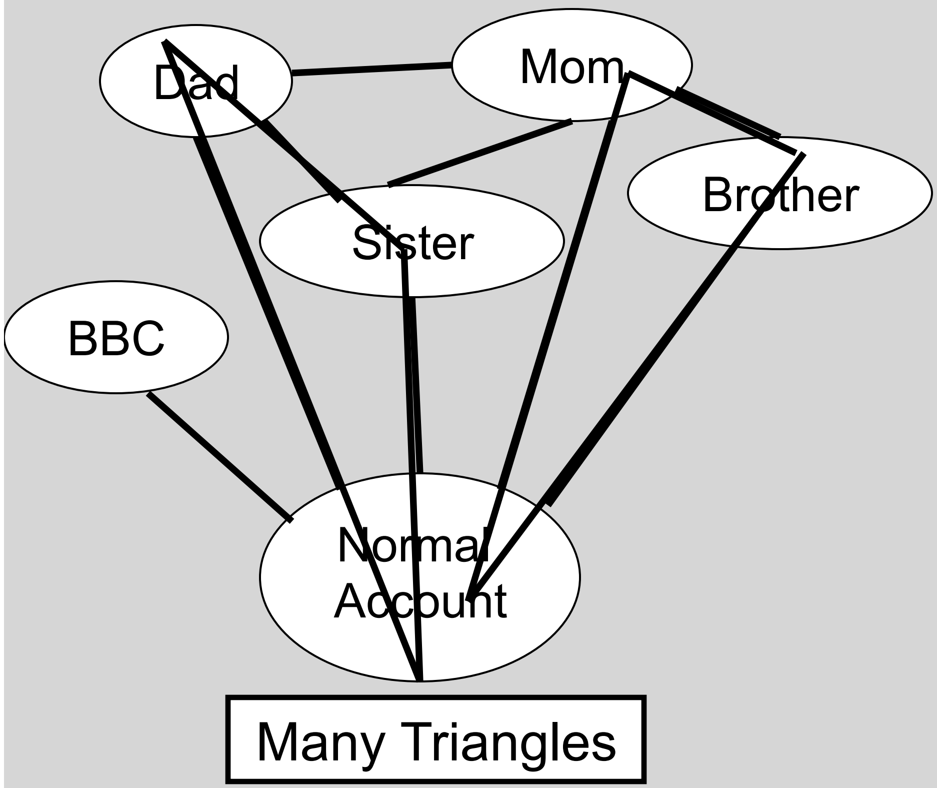
Designing New and Robust Features

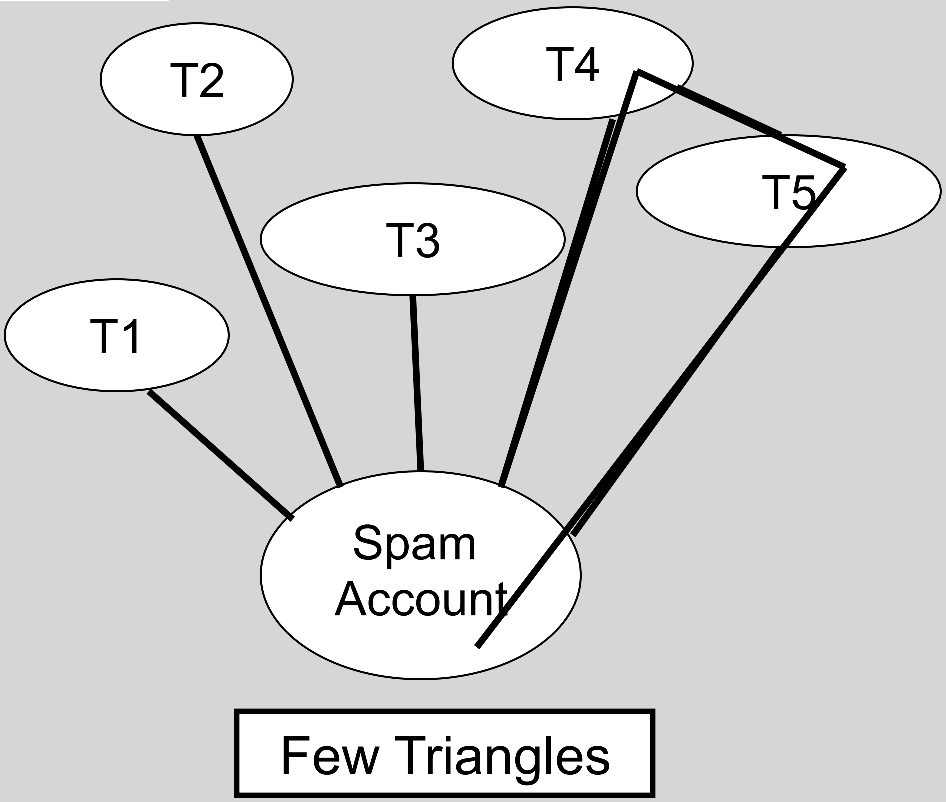
- Graph-Based Features
- Neighbor-based Features
- Automation-based Features
- Timing-based Features

Graph-Based Features

- Local Clustering Coefficient:

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

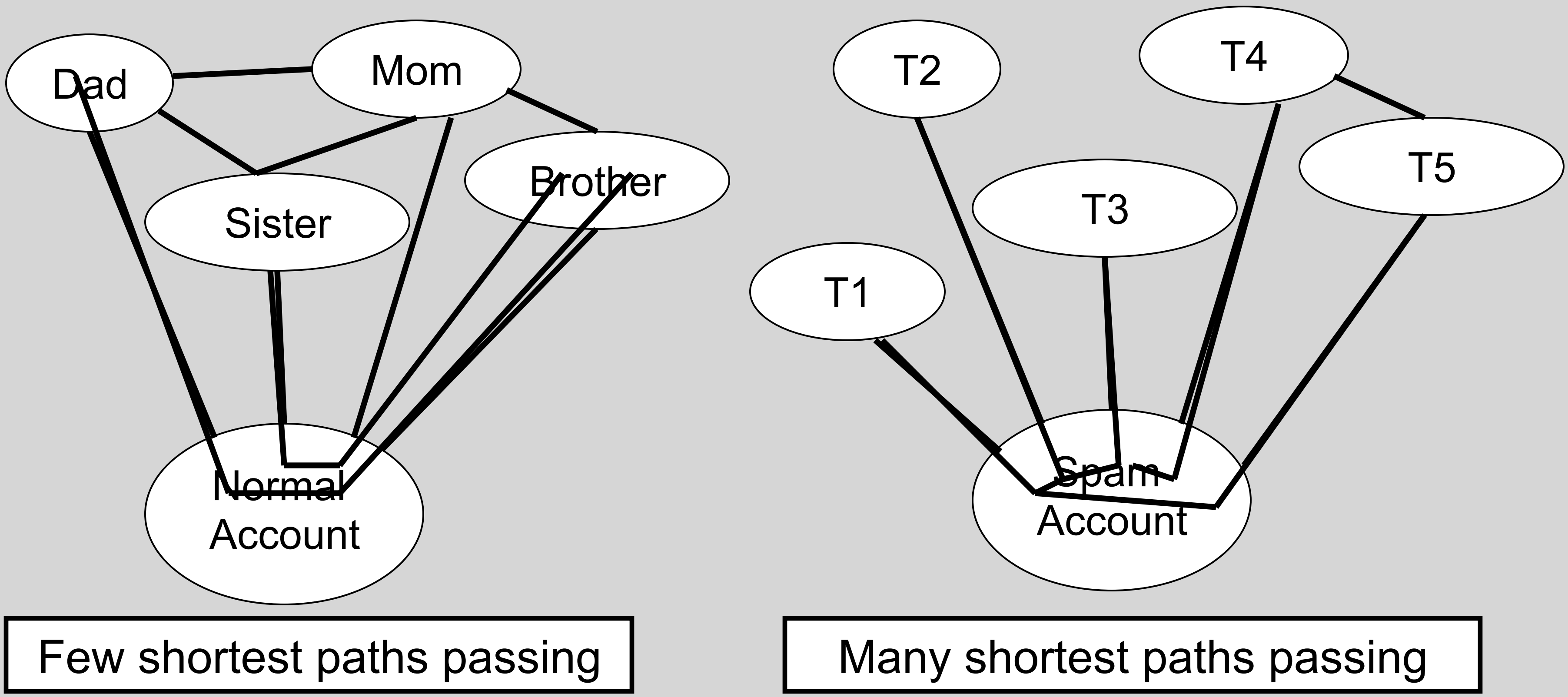




Graph-Based Features

- Betweeness Centrality:

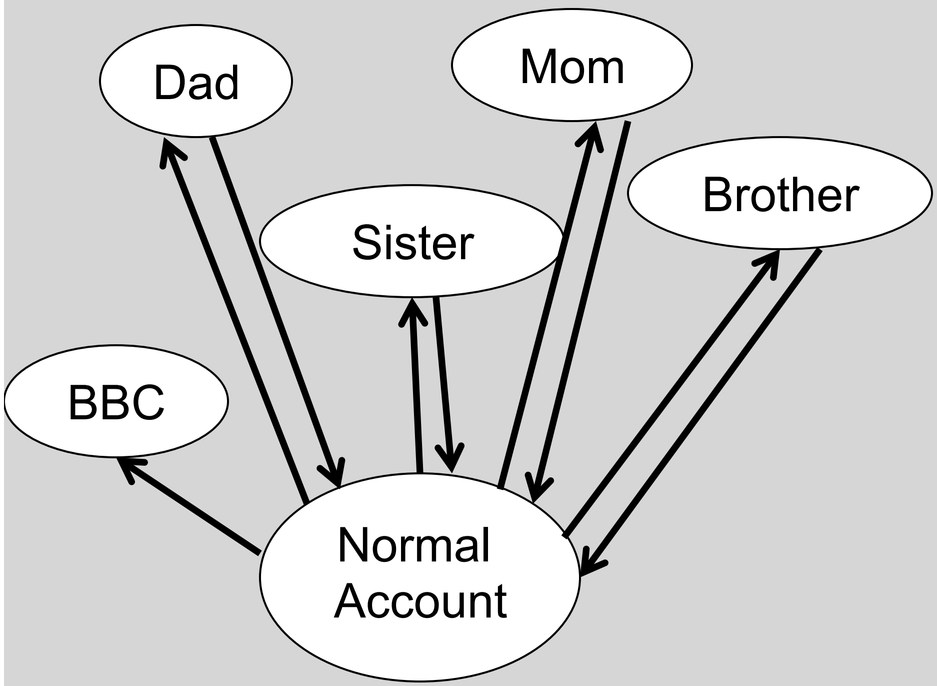
$$BC(v) = \frac{1}{(n-1)(n-2)} \cdot \sum_{s \neq v \neq t \in V} \frac{\delta_{st}(v)}{\delta_{st}}$$



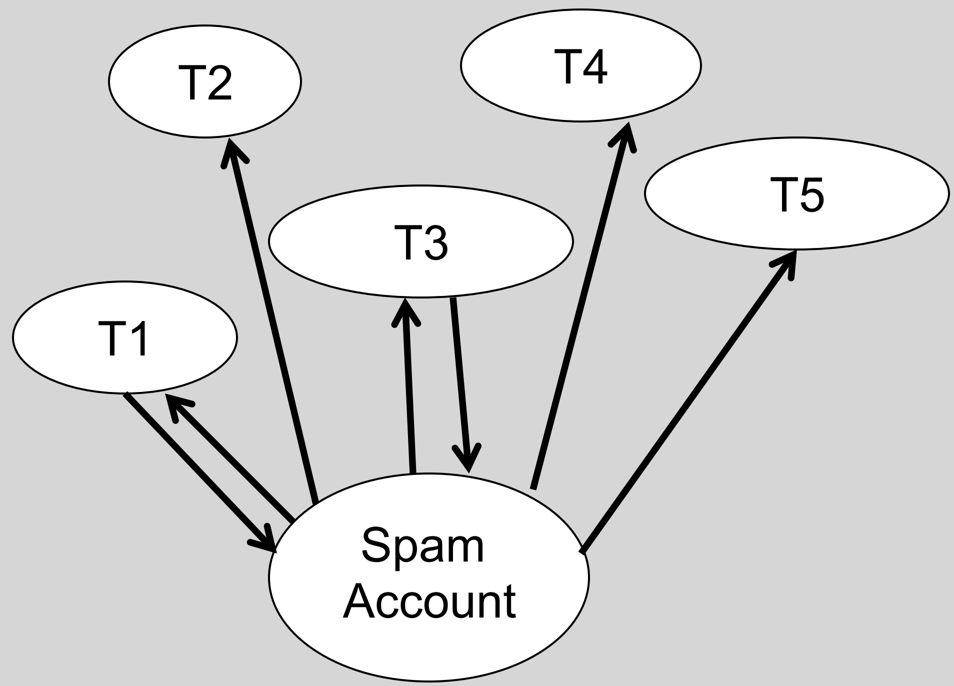
Graph-Based Features

- Bi-directional Links Ratio:

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



Ratio = 4/5 = 80%

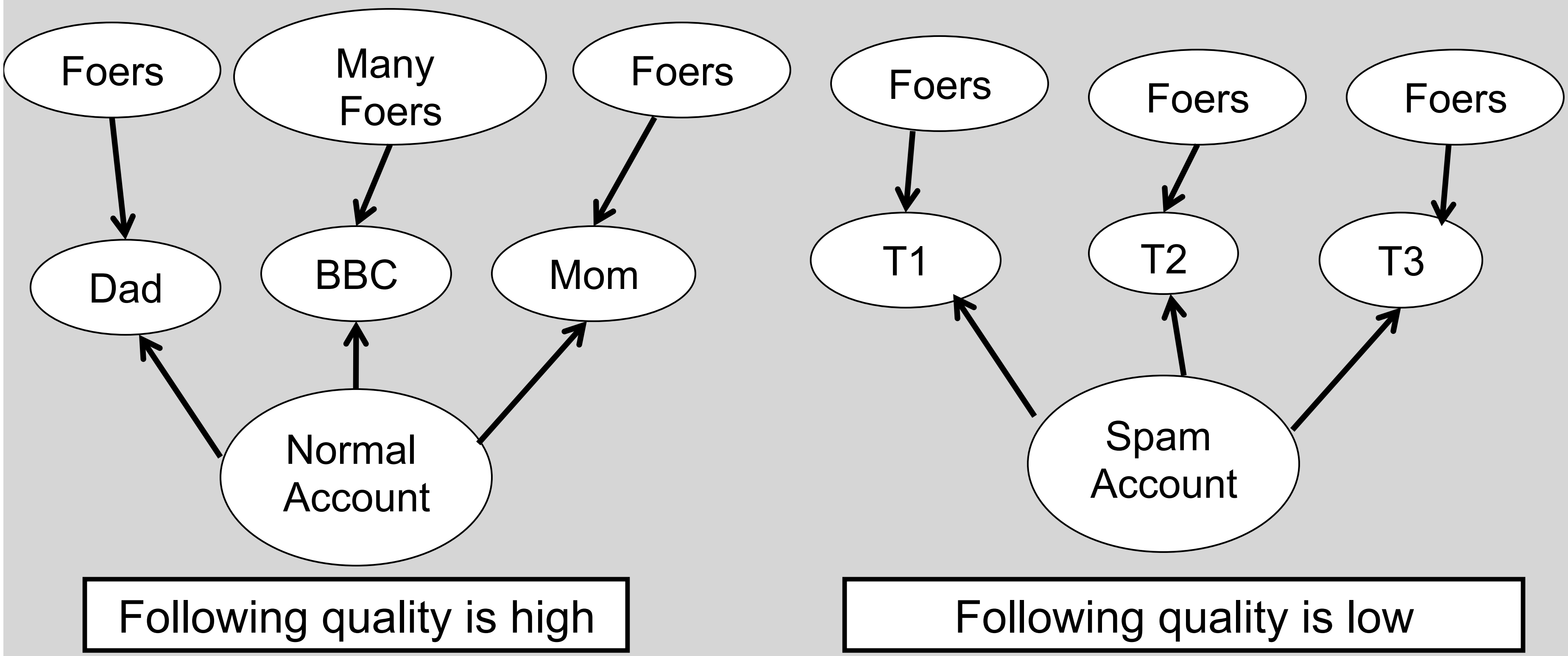


Ratio = 2/5 = 40%

Neighbor-Based Features

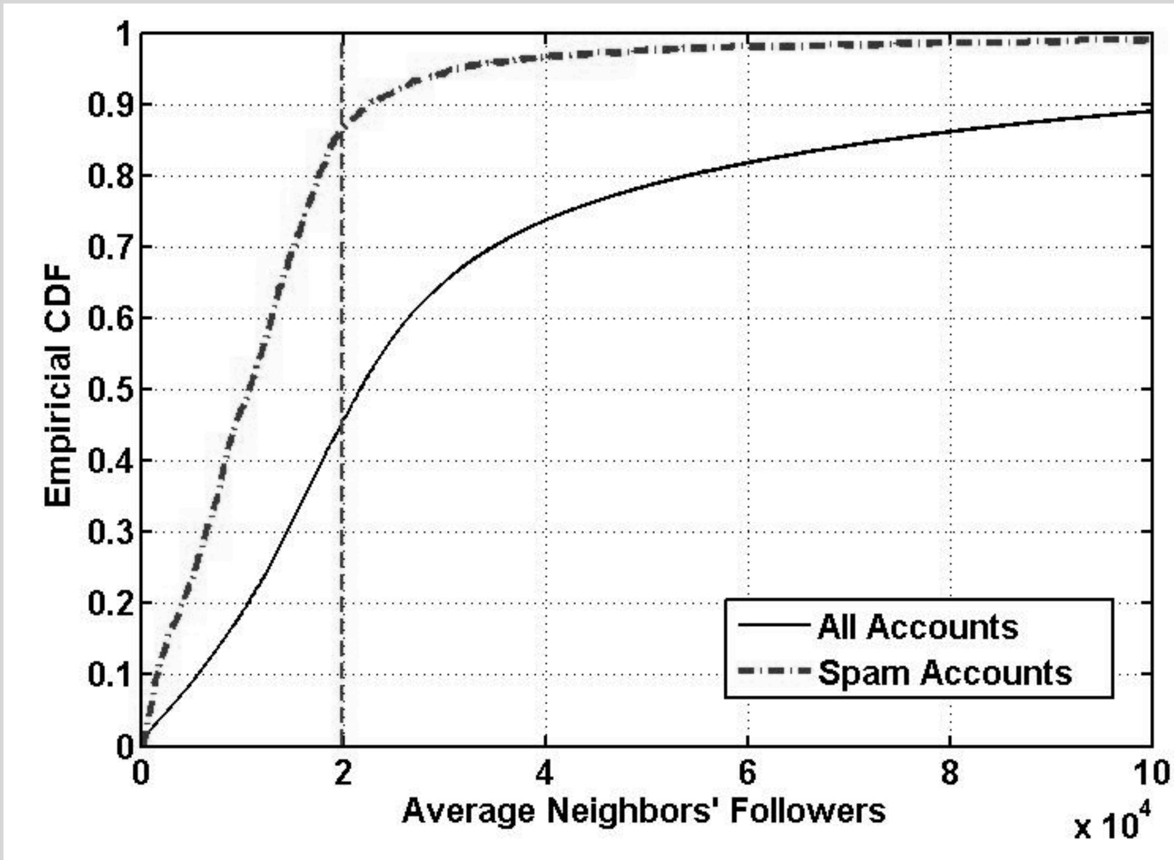
- Average Neighbors' Followers:

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



Neighbor-Based Features

- Average Neighbors' Followers



- 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

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

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

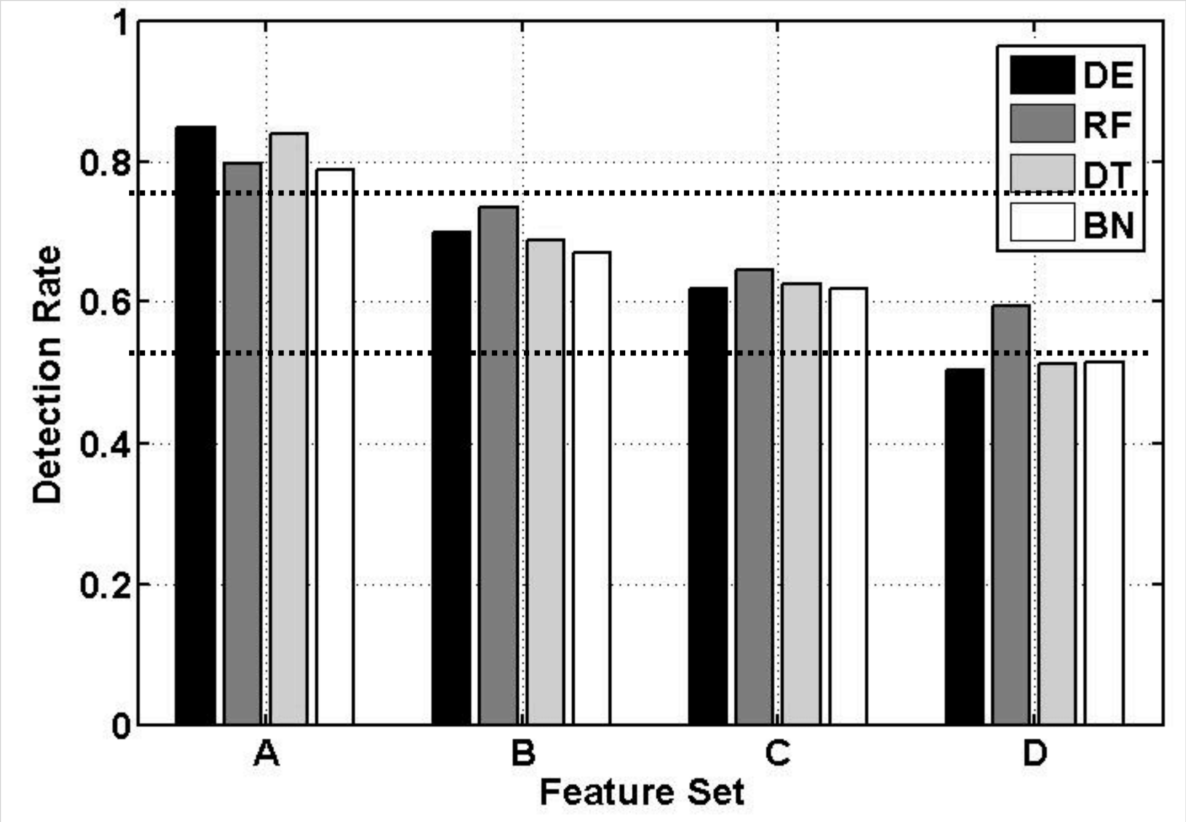
- Similar conclusions can be drawn for the features such as "# of followers" and "following to follower ratio".

Evaluation

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

Performance Comparison

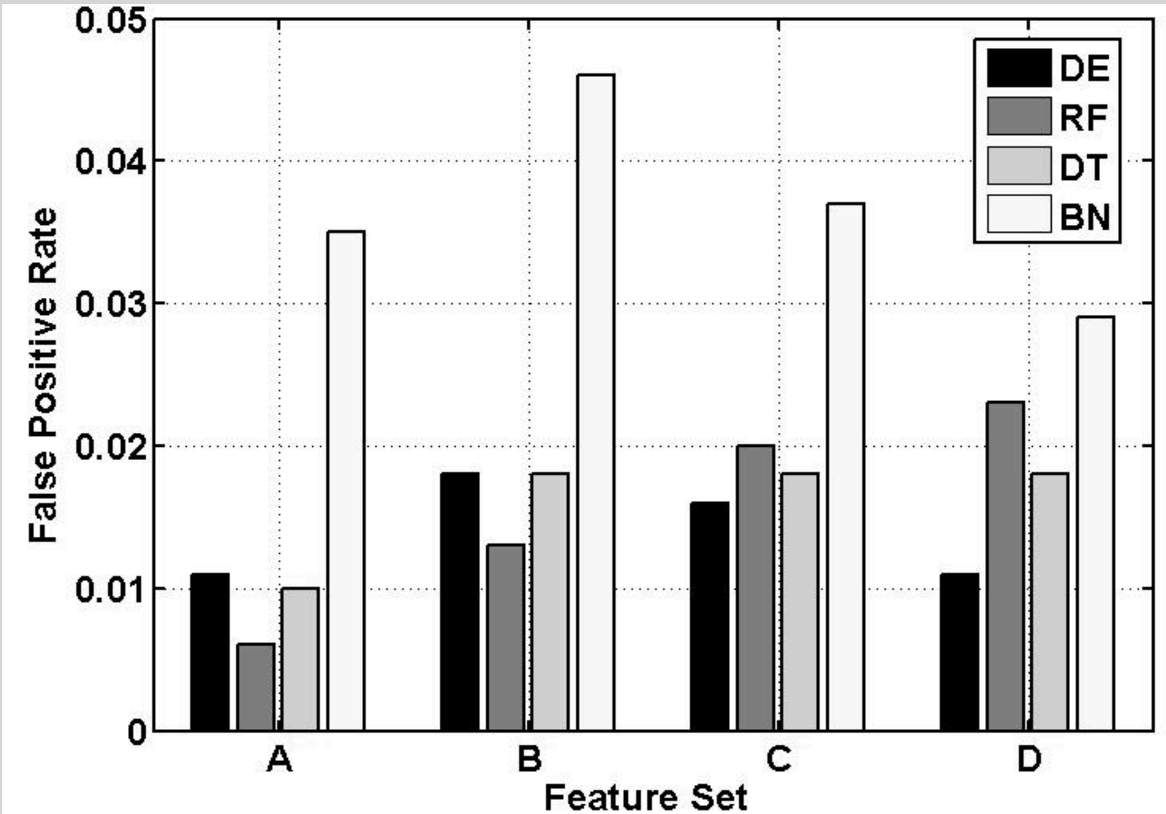
Detection Rate



- 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

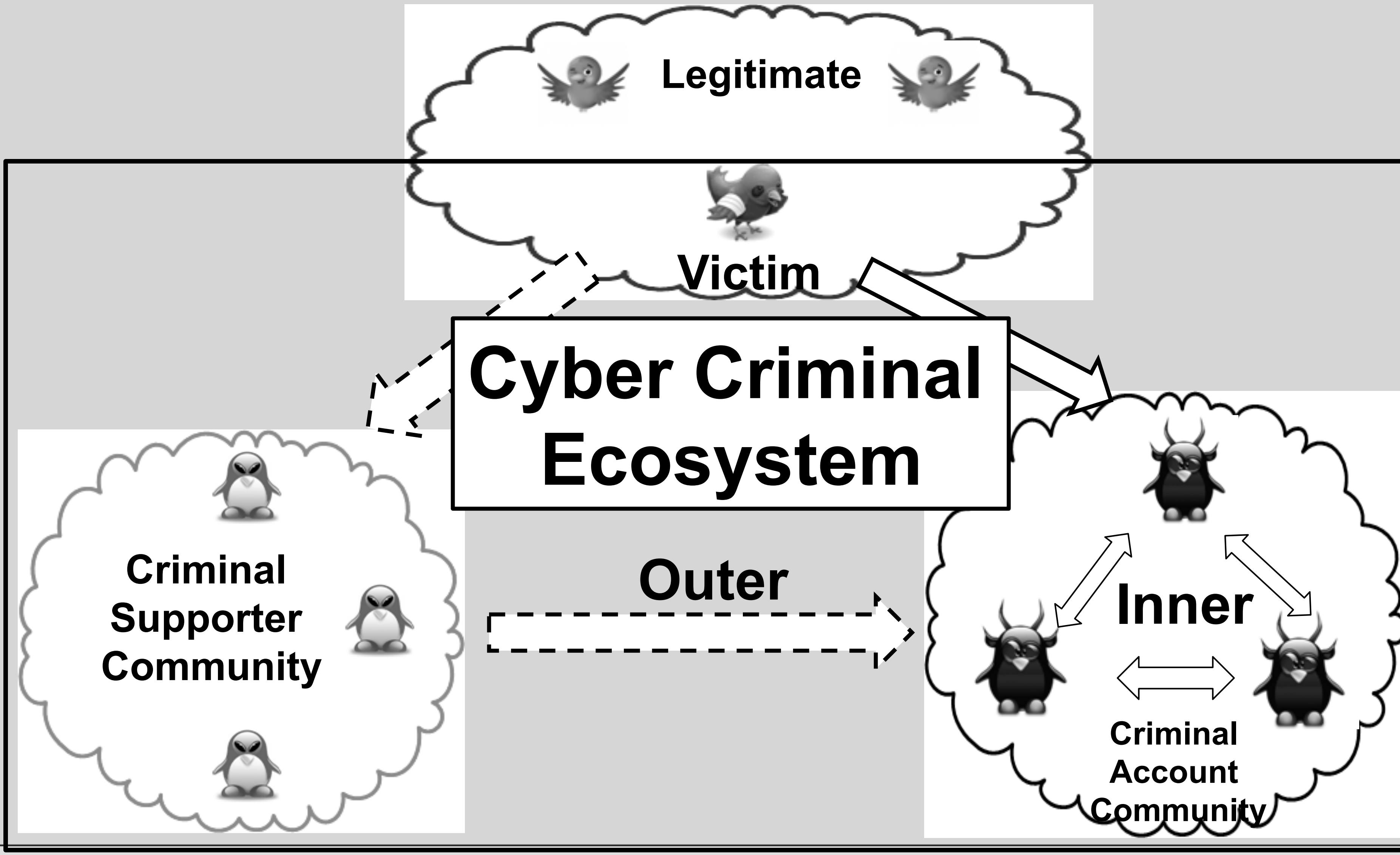
- Newly crawl 3,500 unlabeled accounts
- Used the detector trained on the first data set and use Decorate to classify
- Bayesian detection rate of 88.6% (62/70), 17 accounts post malicious URLs detected by Google Safe Browsing blacklist

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



TEXAS A&M UNIVERSITY

Research Goals

We try to Answer

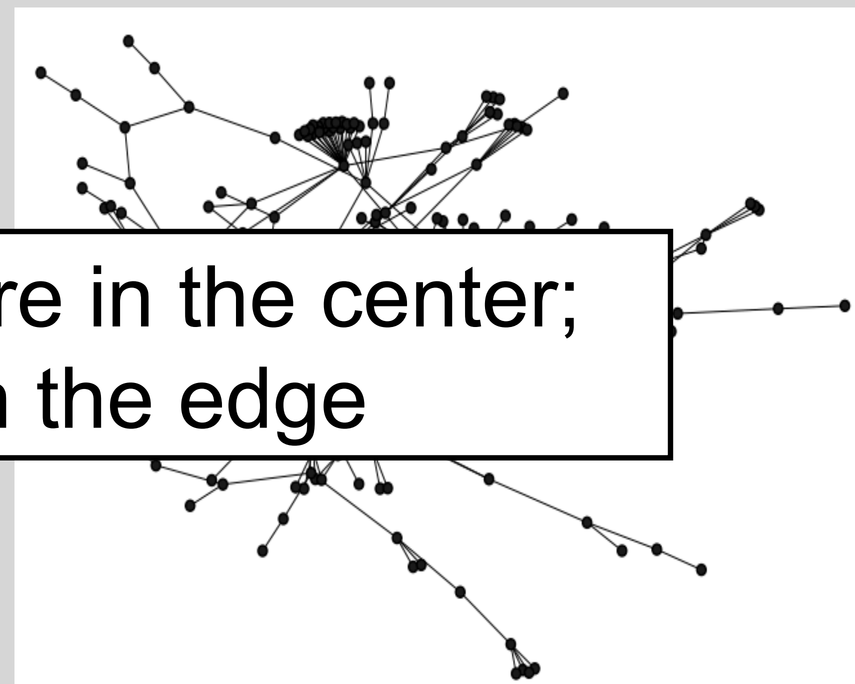
- What is the structure of criminal accounts' network?
- What are possible factors and reasons leading to that structure?
- What are typical characteristics of Criminal Supporters?
- Can we design New Defense Algorithms to Catch more criminal Accounts?
- And SO ON ...

Inner Relationships: Visualizing Relationship Graph

- Node: each criminal account
- Edg

Criminal accounts tend to be socially connected

Relat **nt**



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: $\frac{|E|}{|V| \bullet (|V|-1)}$
    - Criminal graph: 2.33×10^{-3}
    - Public Twitter snapshot (41.7m nodes and 1.47b edges): 8.45×10^{-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

Explainations

- Criminal accounts tend to follow many other accounts without considering those accounts' quality much, making themselves to connect to other criminal accounts.

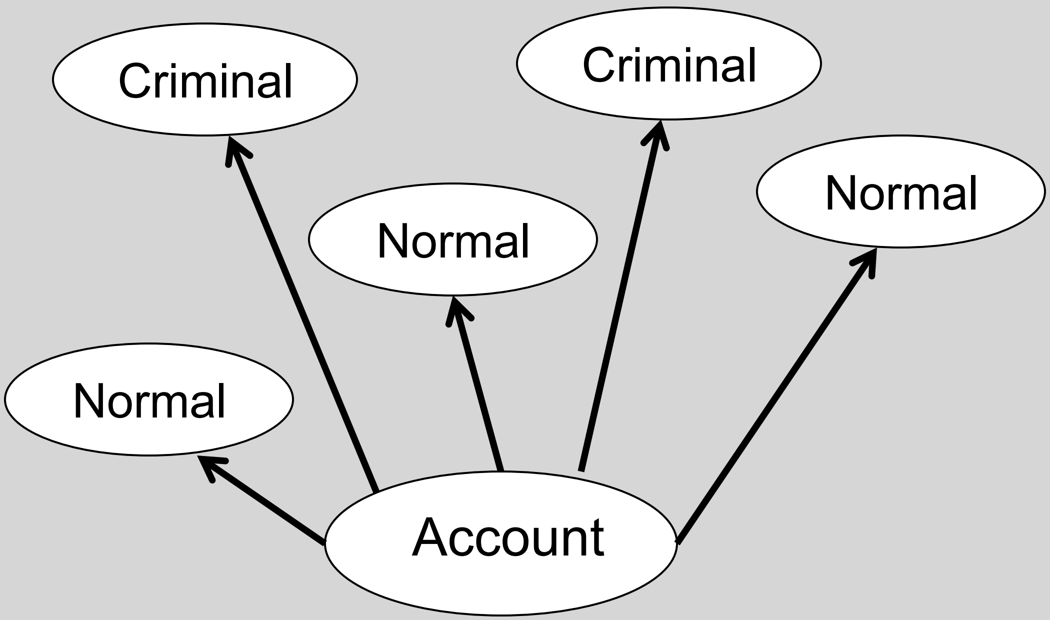
- Criminal accounts, belonging to the same criminal organizations, may be artificially/intentionally connected with each other.

Inner Relationships: Revealing Relationship Characteristics

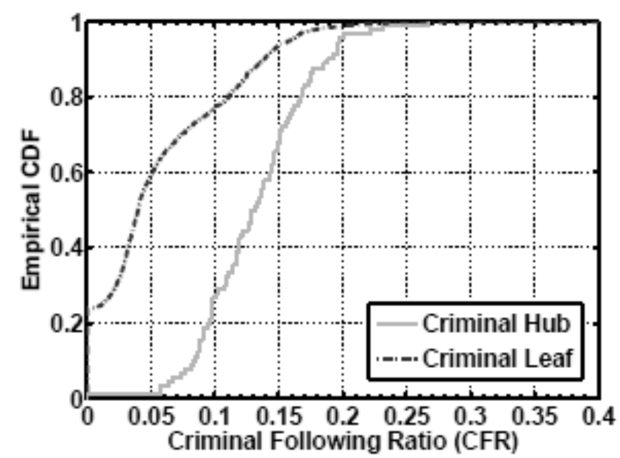
- Observation 2: Compared with criminal leaves (nodes at the edge), criminal hubs (nodes in the center) are more inclined to follow criminal accounts.
  - Extract hubs and leaves: HITS algorithm
  - K-means: 90 hubs, 1970 leaves

Cont.

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



Ratio = 2/5 = 40%



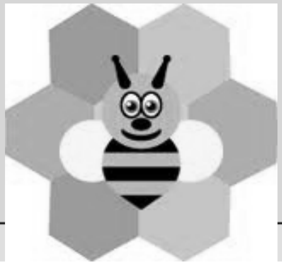
(a) Criminal Following Ratio

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

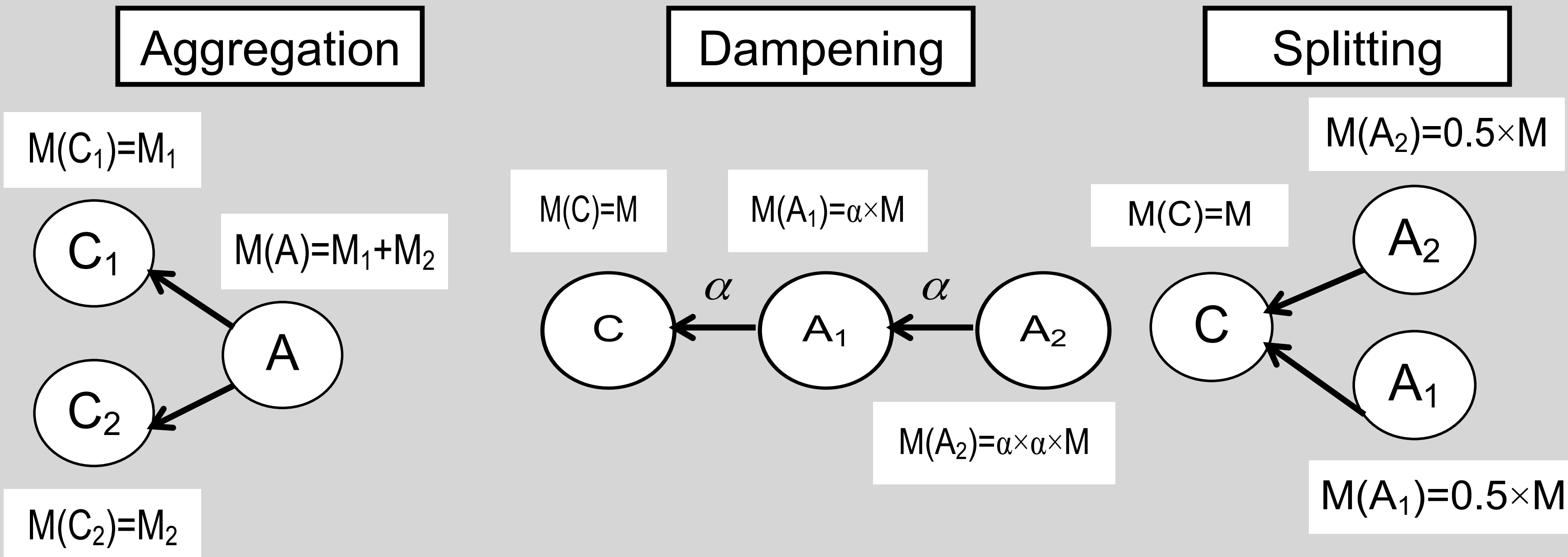
- If criminal accounts mainly build inner social relationships within themselves, criminal accounts can be easily detected.
- However, Twitter criminal accounts have already utilize several tricks to obtain followers outside the criminal account community and mix well into the whole Twitter space.
-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

Outer Social Relationships: Extracting Criminal Supporters

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



- Threshold: x-means; 5,924 criminal

Outer Social Relationships: Characterizing Criminal Supporters

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

Cont.

- Experiment: examine follow backs
- Create 30 Twitter accounts without any tweets and default registration information
- 10
- 10
- 10
- Time span: 48 hours
- Result:
  - Butterflies: 47.8%
  - Normal: 1.8%
  - Criminal: 0.6%

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

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**

Outer Social Relationships: Characterizing Criminal Supporters

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

Cont.

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

  - 6 accounts begin posting (verified) phishing URLs.

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

Inferring Criminal Accounts: Main Idea

- Intuitions:
  - Criminal accounts tend to be socially connected;
  - Criminal accounts usually share similar topics (or keywords or URLs) to attract victims, thus having strong semantic coordination among them.
- 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

- The closeness of social relationships
  - Mr. SPA
- The strength of semantic coordination
  - Semantic Similarity score
  - A higher score between two accounts implies that they have stronger semantic coordination
- Infer criminal accounts in a set of Twitter accounts by starting from a known seed set of criminal accounts
- 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

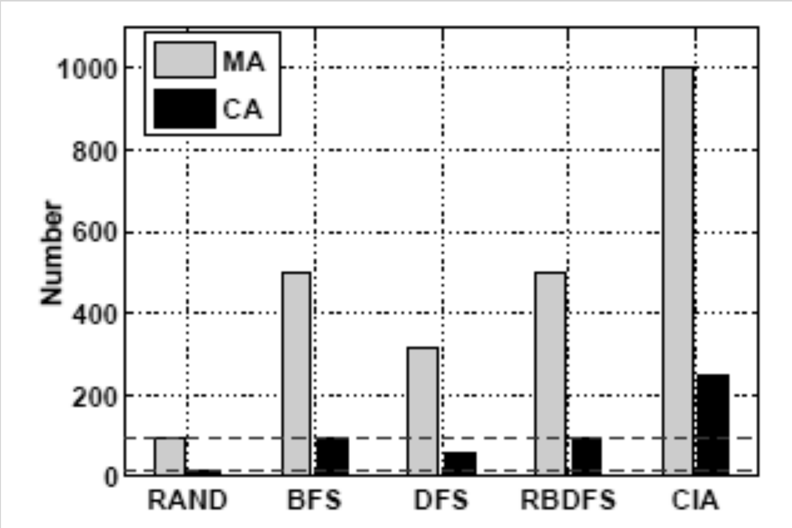
Seed Size = 100

RAND: Randomly Select ;

BFS: Breath First Search

DFS: Depth First Search;

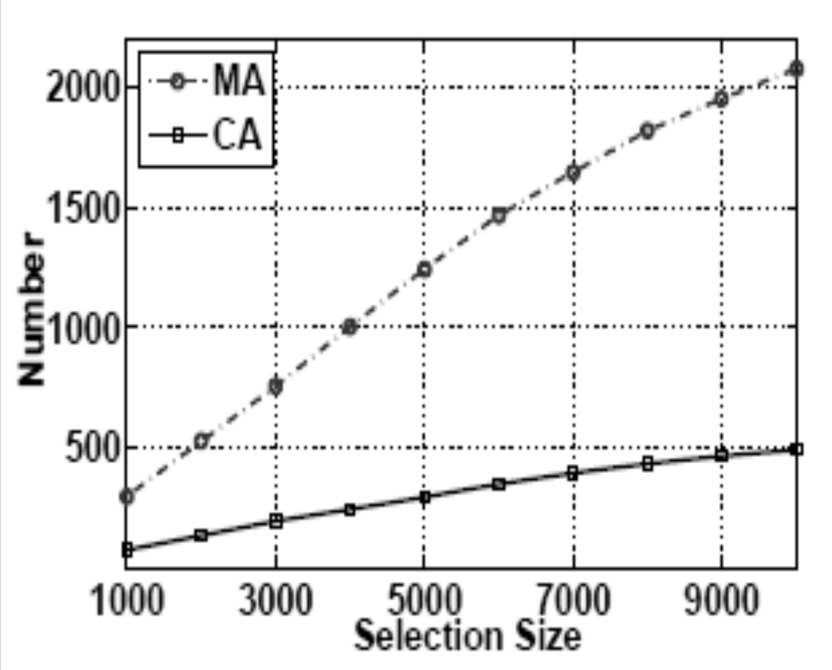
RBDFS: Combine BFS and DFS



Different Selection Sizes

Selection Strategy: CIA

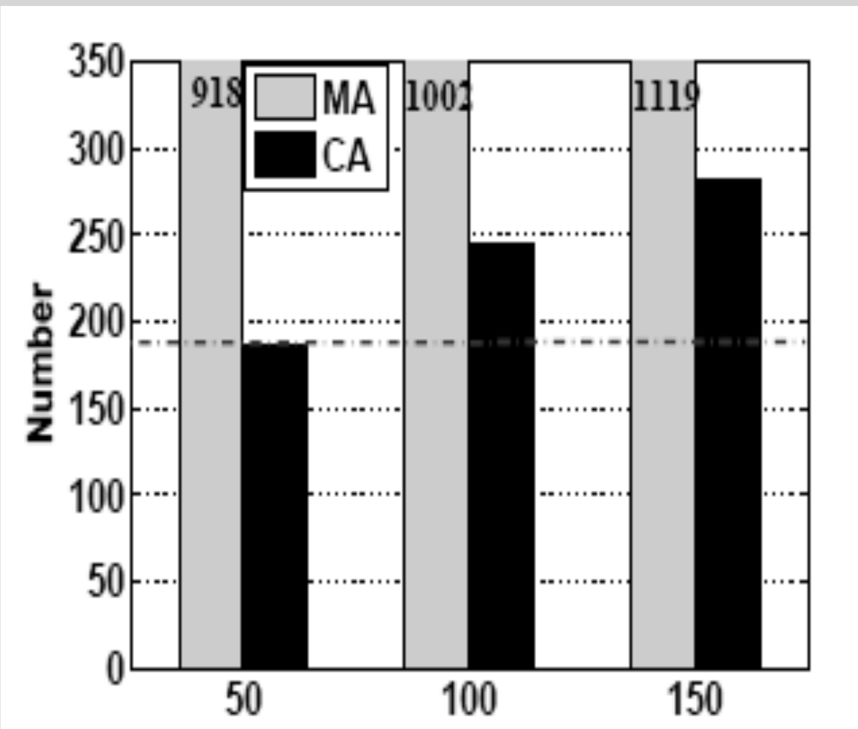
Seed Size = 100



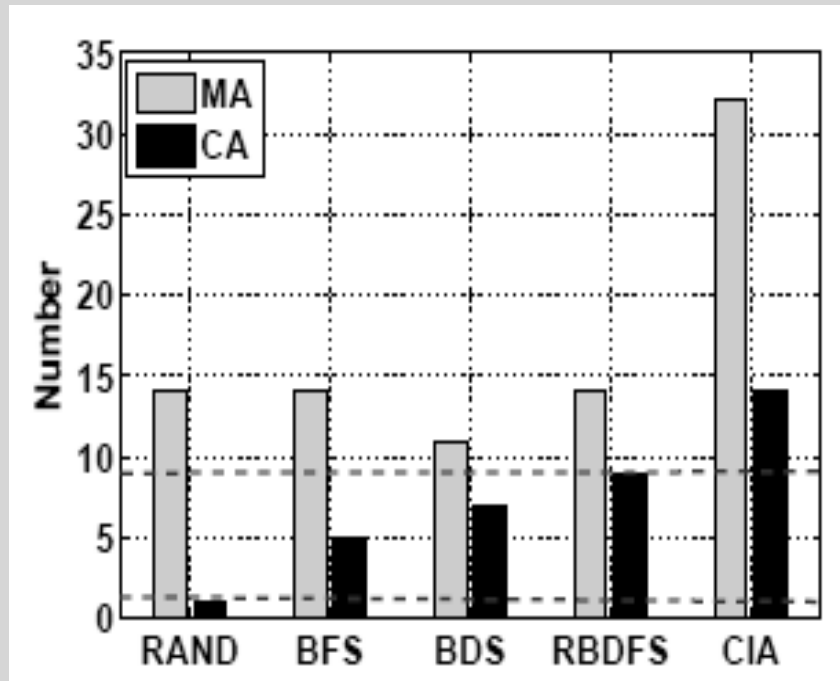
Cont.

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

Limitation

- 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.

- 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.