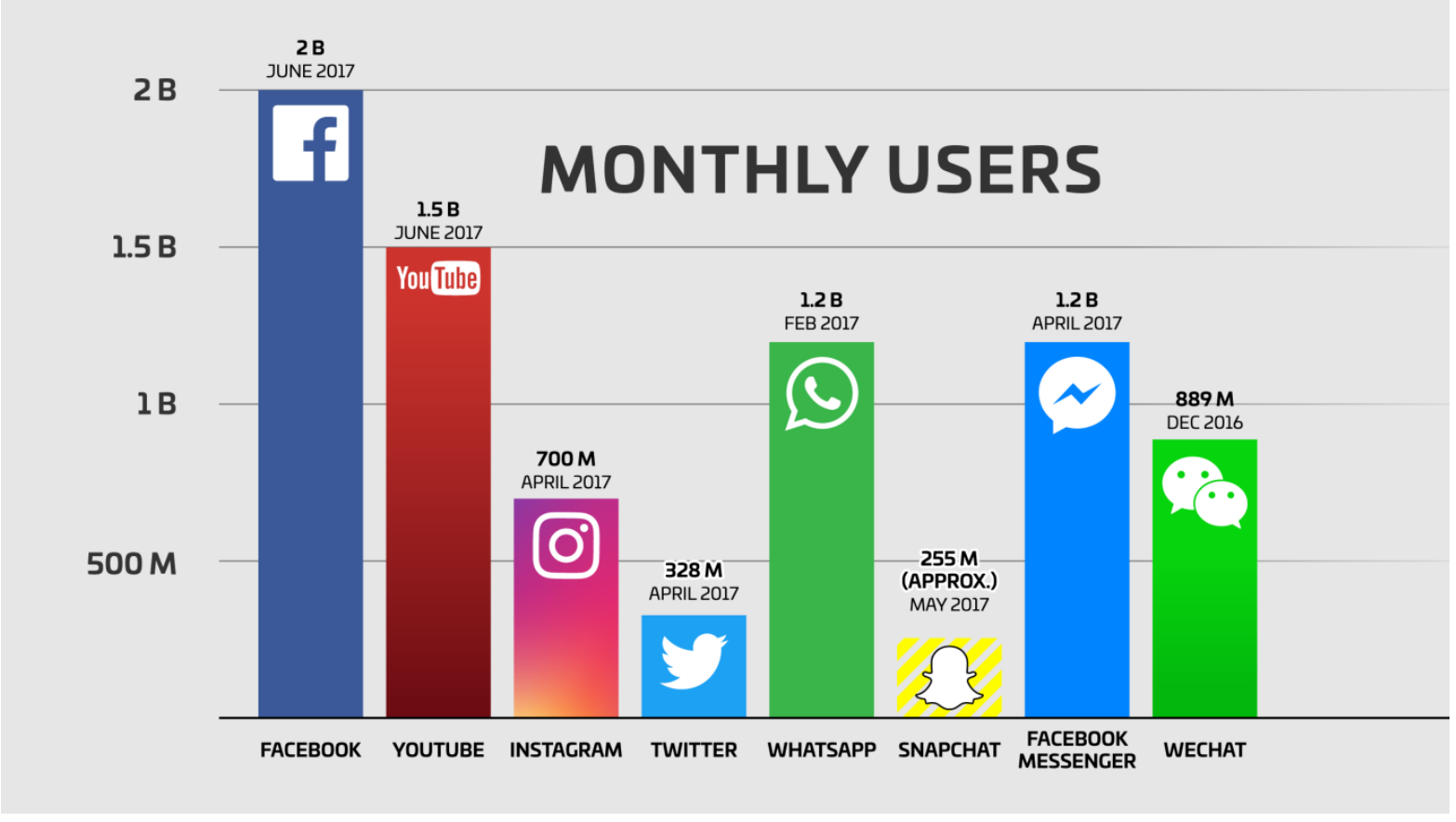
香港中文大學
The Chinese University of Hong Kong

IEEE ICNP 2018

Sybil Detection in Social-Activity Networks: Modelling, Algorithm and Evaluations

Xiaoying Zhang, Xie Hong, John C.S. Lui
Dept. of Computer Science & Engineering
The Chinese University of Hong Kong

Online Social Networks (OSNs)--- Popular & Important



Millions Fake Users (Sybils) exist on OSNs

FACEBOOK SOCIAL

Facebook has disabled almost 1.3 billion fake accounts over the past six months

Facebook will begin publishing more data about how many posts it takes down.

By Kurt Wagner and Rani Molla | May 15, 2018, 10:00am EDT

Technology

BBC NEWS

Twitter 'shuts down millions of fake accounts'

9 July 2018

Share

Threats of Sybils

TECHNOLOGY

How Twitter Bots Are Shaping the Election

Between the first two presidential debates, a third of pro-Trump tweets and nearly a fifth of pro-Clinton tweets came from automated accounts.

DOUGLAS GUILBEAULT AND SAMUEL WOOLLEY NOV 1, 2016

Can social media influence financial markets?

11 Nov 2015

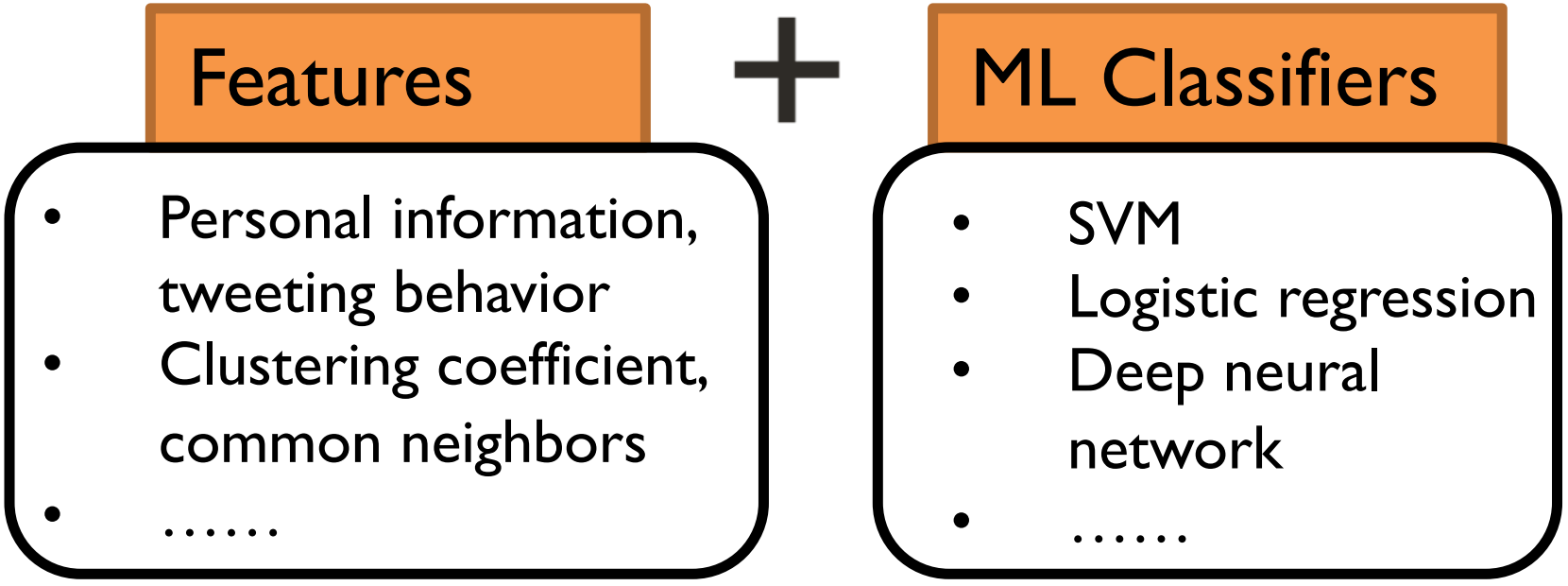
Costas Milas
Professor, University of Liverpool

This article is published in collaboration with The Conversation.

Scottish financial trader James Alan Craig has been charged in the US for allegedly using Twitter to manipulate share prices. According to the US Department of Justice, the 62 year old, from Dunragit in Dumfries and Galloway, caused shareholders to lose more than $1.6m (£1.1m) after allegedly spreading "fraudulent" information about companies on the social network. According to

Existed Sybil detection methods

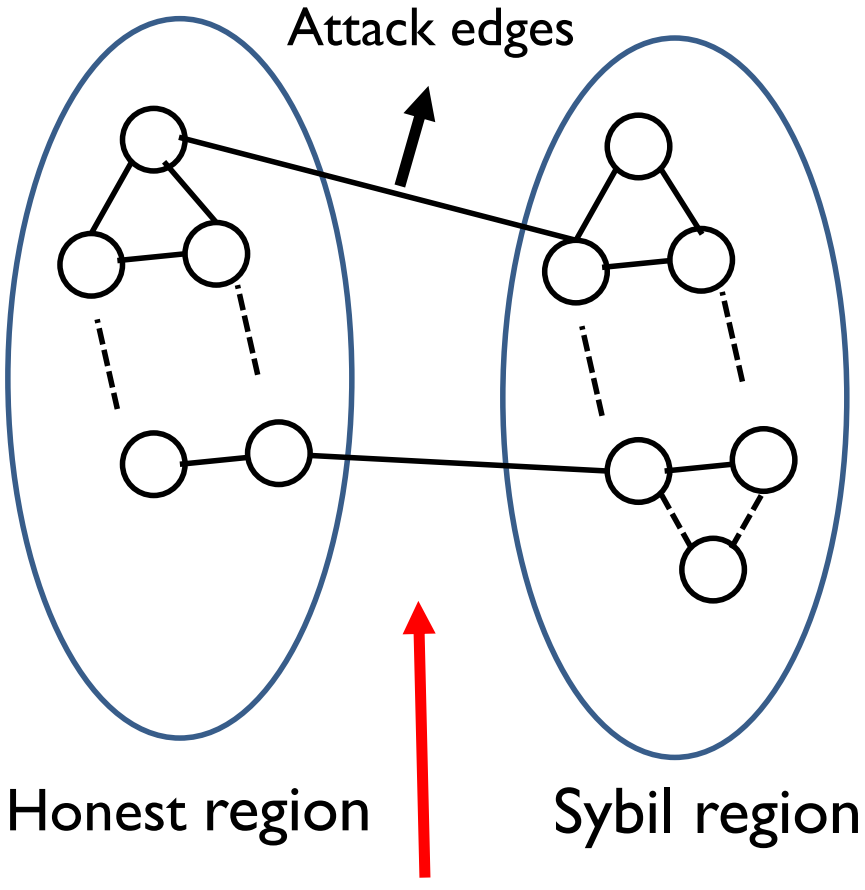
- **Feature-based methods**



Detect only sybils with known patterns !

Existed Sybil detection methods

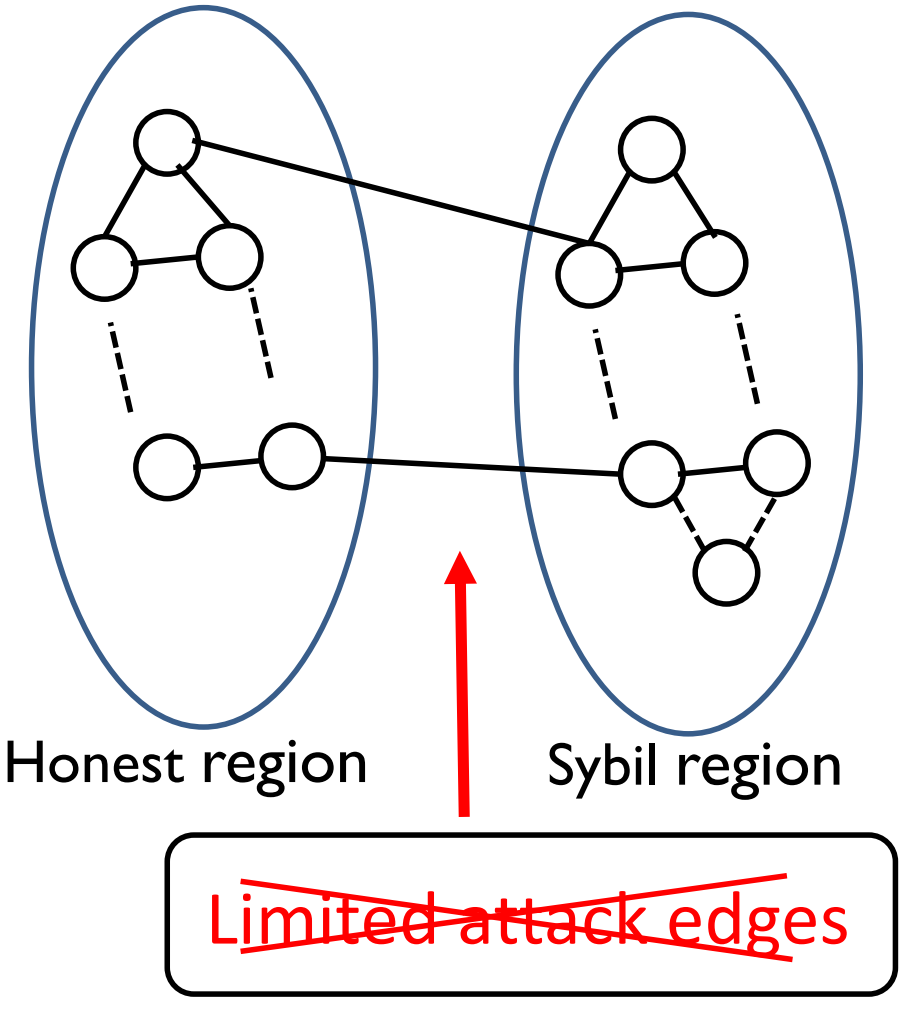
- **Graph-based methods**



Limited-attack-edge assumption: honest users seldom make friends with sybils.

- **Techniques**
  - Random walk (SybilRank [NSDI'12])
  - Fundamental matrix (SybilWalk [DSN'17])
  - Belief propagation (SybilBelief [TIFS'2014],SybilScar[INFCOM' 2017])
  - …….

Inefficiency of Previous Attack Model



- **Limited-attack-edge assumption may not hold !**
  - Sybils can easily befriend with honest users, thus large attack edges exist (Twitter games[ACSAC'14], LinkFarm[WWW'12]).
- Breaking the assumption leads to low accuracy.

Our objective

- What's the realistic attack model?
- How to design efficient algorithms to detect sybils under realistic attack model?

Contributions

Realistic Attack Model

- The Social-Activity Attack Model

Efficient Detection Algorithm

- Sybil_SAN

Theoretical Analysis

- Convergence &Sensitivity analysis

Extensive Experiments

- Synthetic & real datasets from Twitter

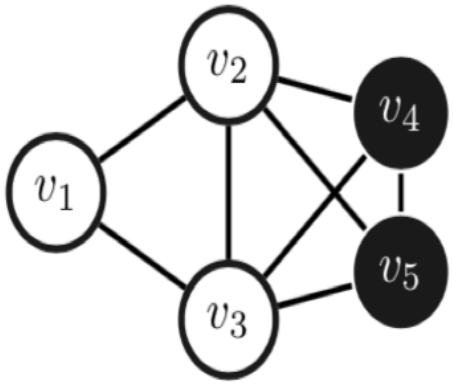
Social-Activity Network Model: Main intuition

Even honest users befriend with sybils, they seldom initiate activities to sybils.

- mention sybils in their own tweets
- reply/retweet sybils' tweets
- ….

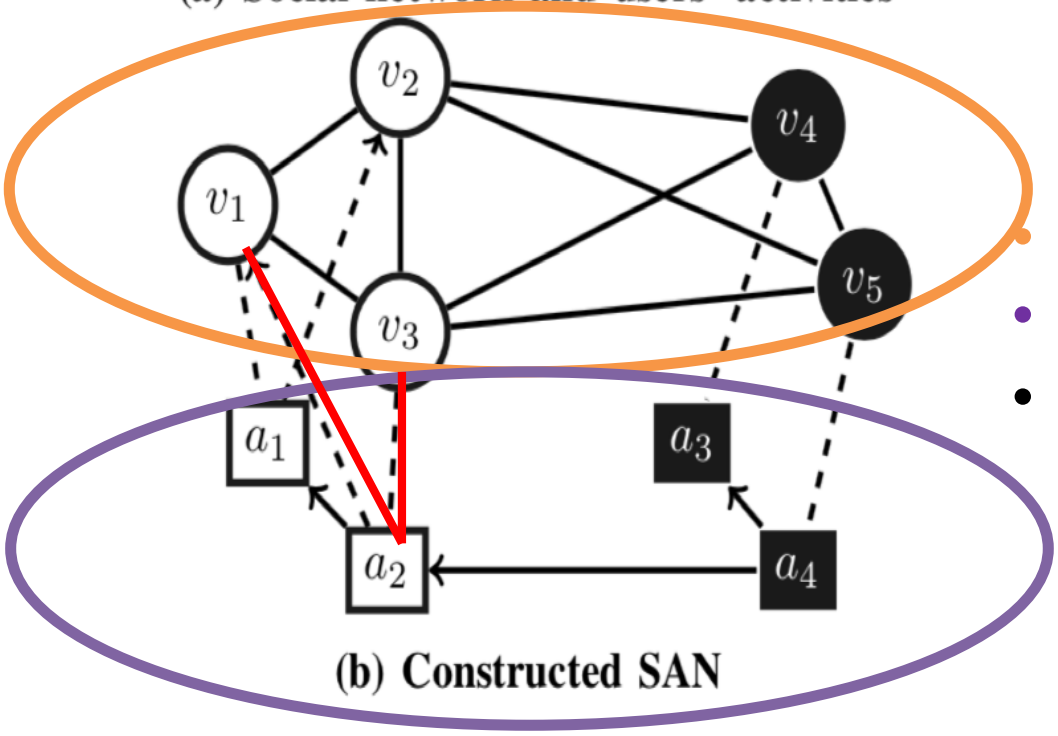
It has been verified by the analysis of a dataset containing thousands of real sybils in Twitter by Zhang et.al[TON'2016].

The Social and Activity Network Model



(v_1, a_1): user v_1 creates tweet a_1
(v_3, a_2): user v_3 creates tweet a_2
(v_4, a_3): user v_4 creates tweet a_3
(v_5, a_4): user v_5 creates tweet a_4
(a_2, a_1): tweet a_2 retweets tweet a_1
(a_4, a_2): tweet a_4 retweets tweet a_2
(a_4, a_3): tweet a_4 replies tweet a_3
(a_2, v_1): tweet a_2 mentions user v_1
(a_1, v_2): tweet a_1 mentions user v_2

(a) Social network and users' activities



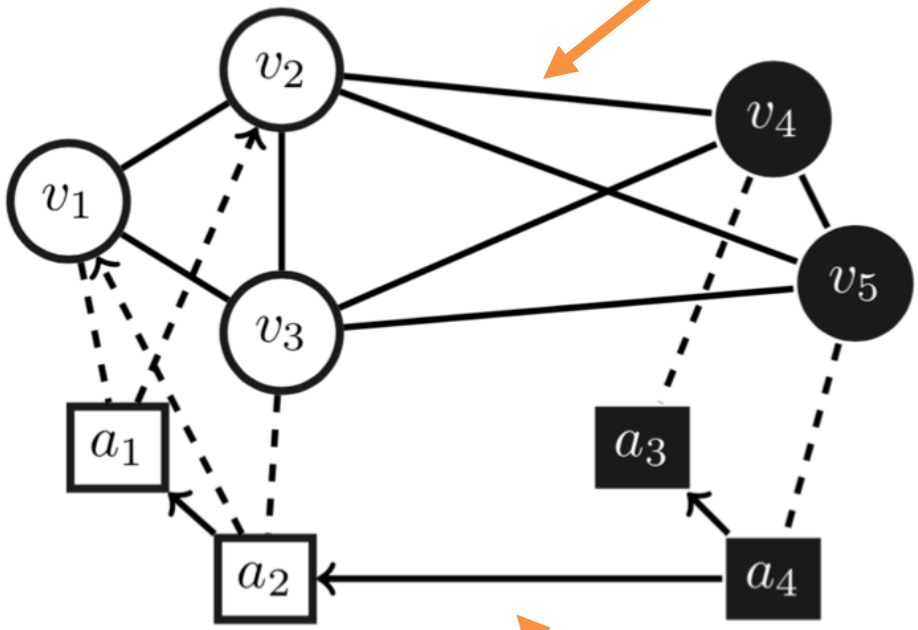
(b) Constructed SAN

- Layer 1: friendship graph G
- Layer 2: Activity graph $\tilde{G}$.
- Between layers: user-activity mapping graph

A More Realistic Attack Model

Friendship attacks (N_A):

- *Property* 1. N_A *can take any value in* $\{0,1,\ldots,|V_h|\times|V_s|\}$.



Honest region

Sybil region

Incoming interactions attacks (αW_h):

- Activities initiated from honest users to sybils
- $\alpha \approx 10^{-5}$ ([TON'2016]).
- W_h: # of activities among honest users

Outgoing interaction attacks (βW_h):

- *Property* 2. β can be arbitrary large.

Contributions

Realistic Attack Model

- The Social-Activity Network Model

Efficient Detection Algorithm

- Sybil_SAN

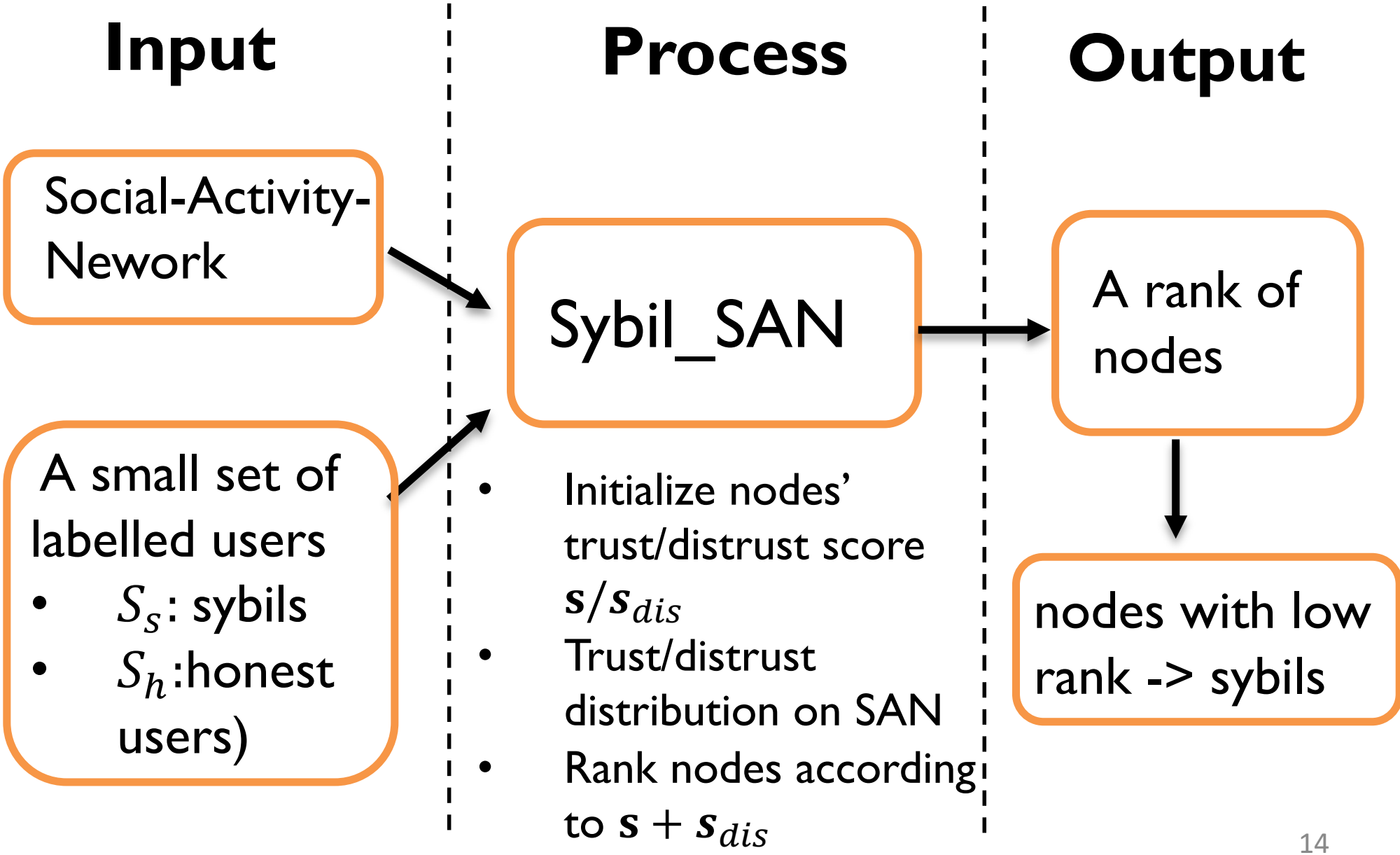
Theoretical Analysis

- Convergence &Sensitivity analysis

Extensive Experiments

- Synthetic & real datasets from Twitter

Sybil_SAN


Input
Process
Output
Social-Activity-Nework
A small set of labelled users
• S_s: sybils
• S_h:honest users)
Sybil_SAN
• Initialize nodes' trust/distrust score $\mathbf{s}/\boldsymbol{s}_{dis}$
• Trust/distrust distribution on SAN
• Rank nodes according to $\mathbf{s} + \boldsymbol{s}_{dis}$
A rank of nodes
nodes with low rank -> sybils

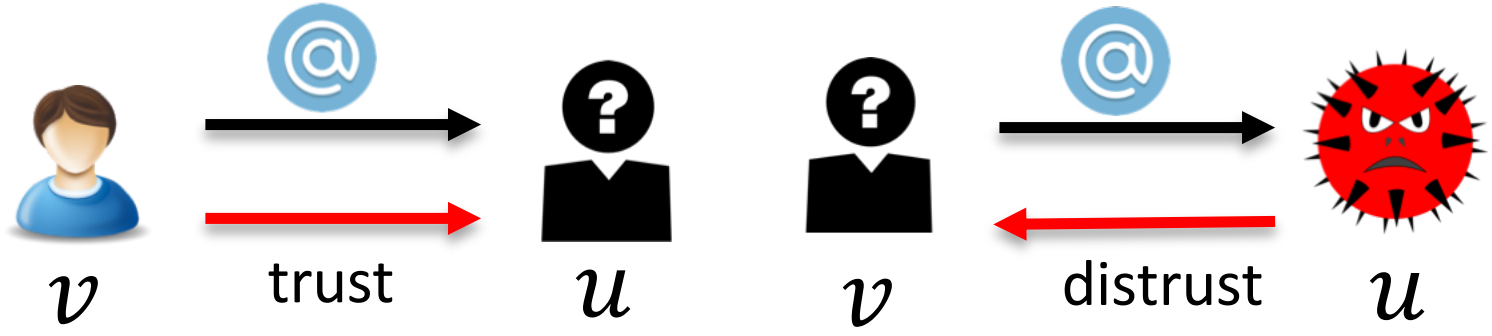
Sybil_SAN

- Initialization $(\mathbf{s}, \boldsymbol{s}_{dis})$

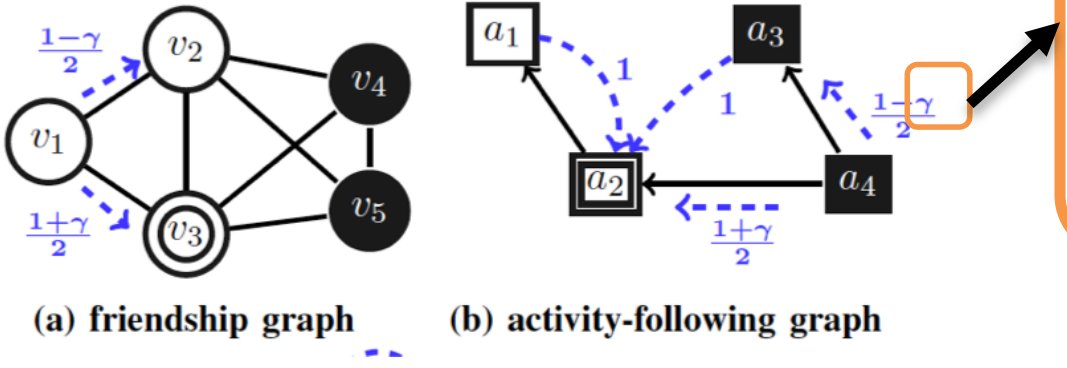
$$\boldsymbol{s_i} = \begin{cases} \frac{\mathbf{1}}{|\boldsymbol{S_h}|}, & \boldsymbol{i} \in \boldsymbol{S_h} \\ \mathbf{0}, & \boldsymbol{o.w} \end{cases} \quad (\boldsymbol{s_{dis}})_{\boldsymbol{i}} = \begin{cases} -\frac{\mathbf{1}}{|\boldsymbol{S_s}|}, & \boldsymbol{i} \in \boldsymbol{S_s} \\ \mathbf{0}, & \boldsymbol{o.w} \end{cases}$$

Sybil_SAN

- Trust/distrust Distribution



  - Distribution in each layer:
    - personalized pagerank (γ)



(a) friendship graph (b) activity-following graph

Tradeoff between exploitation and exploration

Sybil_SAN: Coupled random walks

- **Mutual reinforcement relationship between users and activities.**
  - The activities of a trusted user can be trusted.
  - An activity with high trust score can certify the trustiness of its creator.

Algorithm 2: Coupling random walk

1 The walker starts with node v_i with probability s_i.
2 **repeat**
3 **if** *The walker is at a user node* $v_i \in \mathcal{V}$ **then**
4 With probability λ_i it walks one step on the friendship graph according to the *WalkOnFriendGraph* algorithm.
5 With probability $(1-\lambda_i)$, it walks $2k+1$ steps on the user-activity graph according to the *WalkOnUserActivityGraph* algorithm.
6 **if** *The walker is at an activity node* $a_i \in \mathcal{A}$ **then**
7 With probability $\lambda_{|\mathcal{V}|+i}$, the walker walks n steps on the activity-following graph according to the *WalkOnActivityFollowingGraph* algorithm.
8 With probability $(1-\lambda_{|\mathcal{V}|+i})$ it takes $2k+1$ steps on the user-activity graph according to the *WalkOnUserActivityGraph* algorithm.
9 **until** *converge*

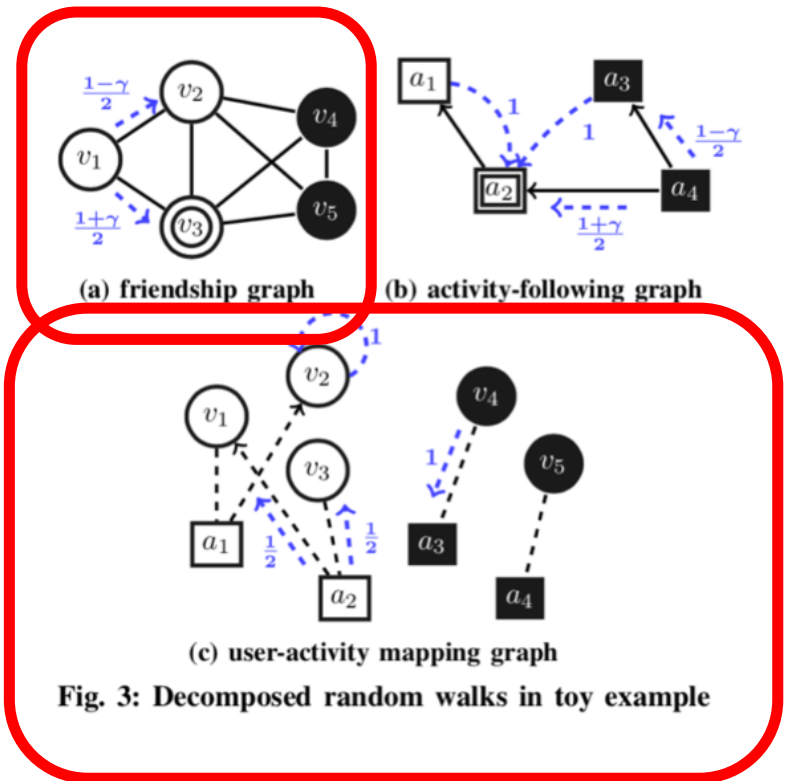


Fig. 3: Decomposed random walks in toy example

Contributions

Realistic Attack Model

- The Social-Activity Network Model

Efficient Detection Algorithm

- Sybil_SAN

Theoretical Analysis

- Convergence &Sensitivity analysis

Extensive Experiments

- Synthetic & real datasets from Twitter

Theoretical Analysis: Convergence

Theorem 1:

Sufficient conditions to guarantee Sybil_SAN **converge** to **unique** trust score

- friendship graph $\mathcal{G}$ is connected
- $0 < \gamma < 1$
- $0 < \lambda_i < 1, \forall i$ satisfying $i > |\mathcal{V}|$ or v_i has activities

Theorem 2:

Suppose $\left|\left|\boldsymbol{s}^{t+1} - \boldsymbol{s}\right|\right| \leq \epsilon$ in Sybil_SAN is measured by $1-$norm, then Sybil_SAN stops in at most $1 + \frac{1}{v}\ln(\frac{4}{\epsilon s^*_{min}})$ rounds, where

- $s^*_{min} = \min_i s^*_i$
- v: spectral gap of the Markov chain

Theoretical Analysis: Sensitivity analysis

Trust score under graph without attack edges

Trust score estimated by Sybil_SAN

Theorem 3:

$$\frac{\left|\left|\boldsymbol{s}^{*}-\tilde{\boldsymbol{s}}^{*}\right|\right|}{\left|\left|\tilde{\boldsymbol{s}}^{*}\right|\right|} \leq \epsilon_{sd}$$

where ϵ_{sd} is defined as :

$$\epsilon_{sd} \triangleq \left|\left|\left[(\boldsymbol{P}_{cr}-\boldsymbol{I})(\boldsymbol{P}_{cr}^{T}-\boldsymbol{I}\,)+\boldsymbol{e}^{T}\boldsymbol{e}\right]^{-1}\times(\boldsymbol{E}(\boldsymbol{P}_{cr}-\boldsymbol{I}+\boldsymbol{E}^{T})+(\boldsymbol{P}_{cr}-\boldsymbol{I})\boldsymbol{E}^{T})\right|\right|$$

Contributions

Realistic Attack Model

- The Social-Activity Network Model

Efficient Detection Algorithm

- Sybil_SAN

Theoretical Analysis

- Convergence &Sensitivity analysis

Extensive Experiments

- Synthetic & real datasets from Twitter

Synthetic datatsets

- **Honest region**:
  - a public Twitter dataset([Weng et.al 2013])

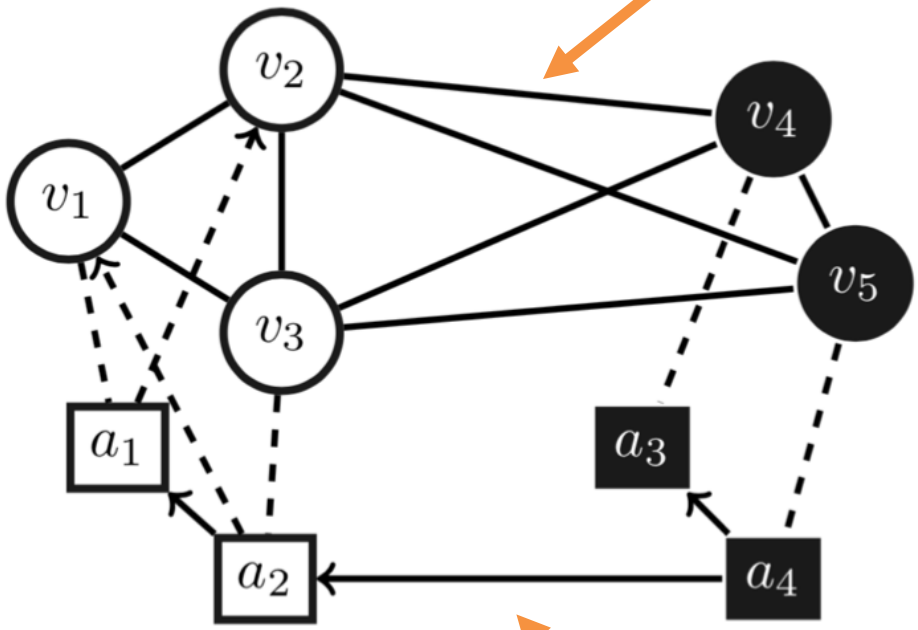
# of nodes	# of edges	# of activities
543,785	28,397,413	21,426,709

  - Two type of activities:
    - user v_i retweets user v_j's tweets
    - user v_i mentions user v_j
- **Sybil region** (N_S, M):
  - M disconnected clusters, all together N_S sybils.
  - the Preferential Attachment (PA) model
- **Attack** (N_A, α, β)

A More Realistic Attack Model

Friendship attacks (N_A):

- *Property* 1. N_A *can take any value in* $\{0,1,\ldots,|V_h|\times|V_s|\}$.



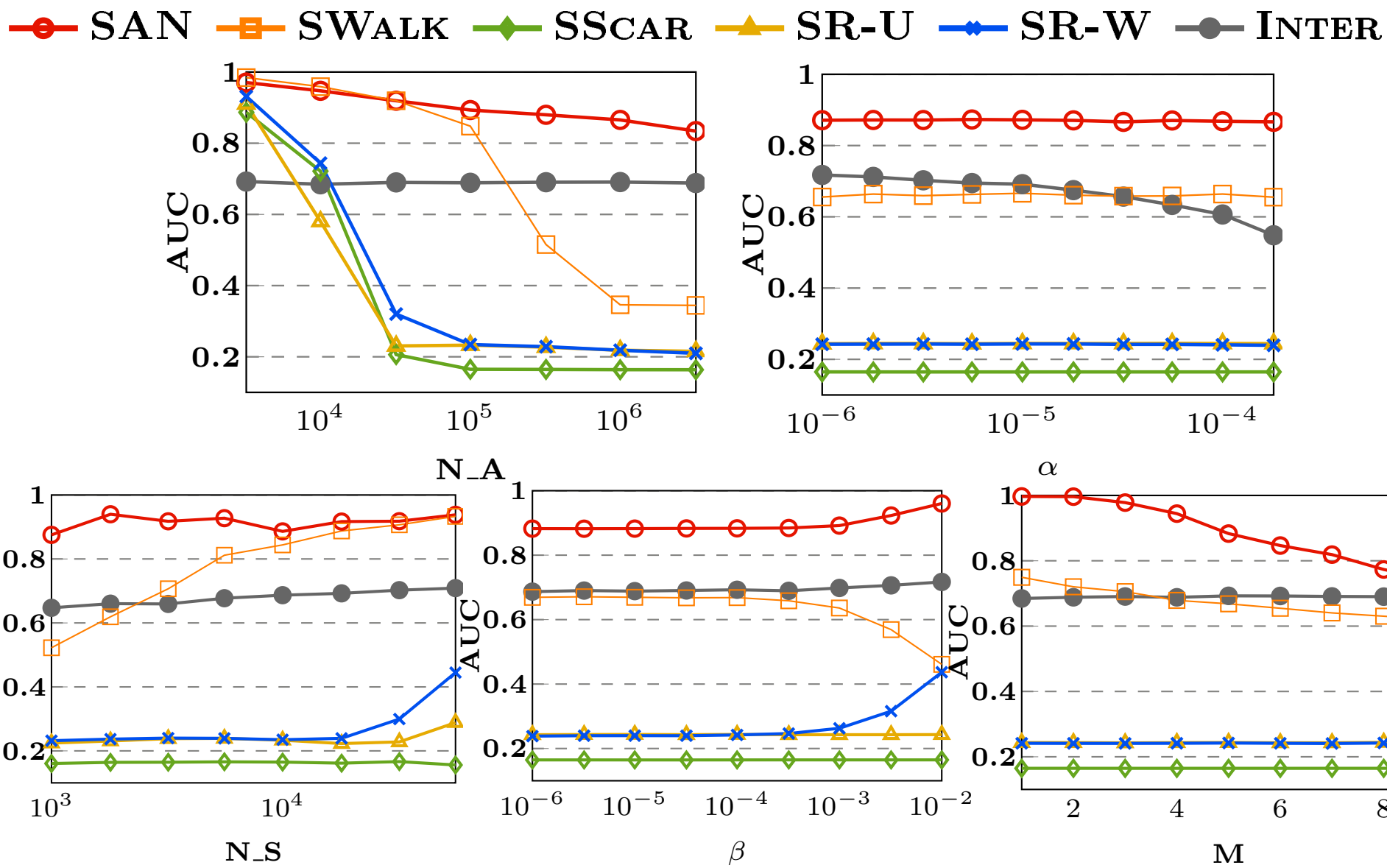
Incoming interactions attacks (αW_h):

- Activities initiated from honest users to sybils
- $\alpha \approx 10^{-5}$ ([TON'2016]).
- W_h: # of activities among honest users

Outgoing interaction attacks (βW_h):

- *Property* 2. β can be arbitrary large.

Experimental results



Sybil_SAN outperforms compared algorithms under various attacks.

Experiments on real dataset

- A crawled subnetwork from Twitter starting from public 991 sybils.
  - Sybils: blocked users
  - Honest users: unblocked users

# of honest users	# of sybils	# of edges	# of interactions
409, 694	40, 548	222,944,310	102,693,769

Experiments on real dataset

- A crawled subnetwork from Twitter starting from public 991 sybils.
  - Sybils: blocked users
  - Honest users: unblocked users (noisy).
- **Results**

	Inter	SR_W	SR_U	SScar	SWalk	SAN
AUC	0.62	0.48	0.52	0.15	0.44	**0.73**
Improved ratio	**17.7%**	**52.1%**	**40.4%**	**386%**	**66%**	

Contributions

Realistic Attack Model

- The Social-Activity Network Model

Efficient Detection Algorithm

- Sybil_SAN

Theoretical Analysis

- Convergence &Sensitivity analysis

Extensive Experiments

- Synthetic & real datasets from Twitter

thank you!