

Sybil Detection and Defense

Yue Duan

Illinois Institute of Technology

Sybil Attacks

- A Sybil attack is a kind of security threat on an online system where **one person tries to take over the network by creating multiple accounts, nodes or computers.**
- Potential problems:
 - out-vote the honest nodes
 - 51% attack
 - block users from the network

SybilGuard: Defending Against Sybil Attacks via Social Networks

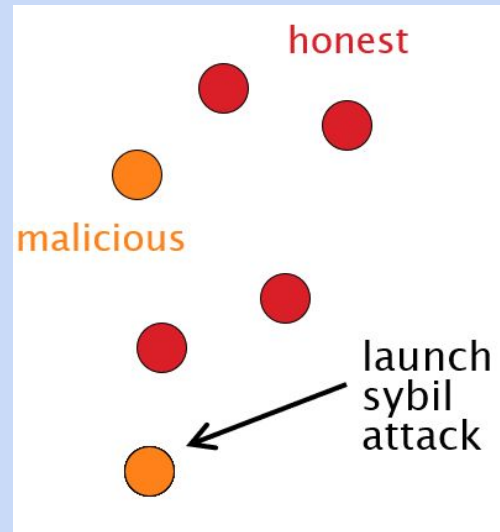
Haifeng Yu* , Michael Kaminsky* , Phillip B. Gibbons* , Abraham Flaxman#

Intel Research*, Carnegie Mellon University#

SIGCOMM 2006

Background

- Sybil attack
 - Single user pretends many fake/sybil identities
 - Creating multiple accounts from different IP addresses
- Sybil identities can become a large fraction of all identities
 - Out-vote honest users in collaborative tasks



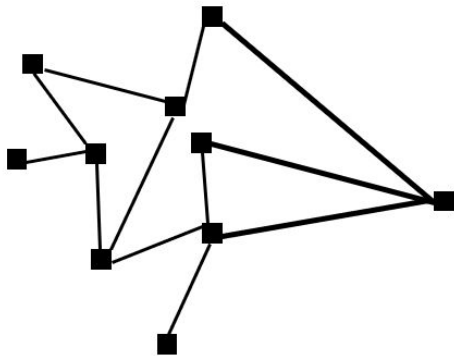
Background

- Defense mechanism
 - Using a trusted central authority
 - Tie identities to actual human beings
 - Not always desirable
 - Can be hard to find such authority
 - Sensitive info may scare away users
 - Potential bottleneck and target of attack
 - Without a trusted central authority
 - Impossible unless using special assumptions [Douceur'02]
 - Resource challenges not sufficient -- adversary can have much more resources than typical user

SybilGuard

- Main Idea: Use a social network as the “central authority”
- A node trusts its neighbors
- Each node learns about the network from its neighbors

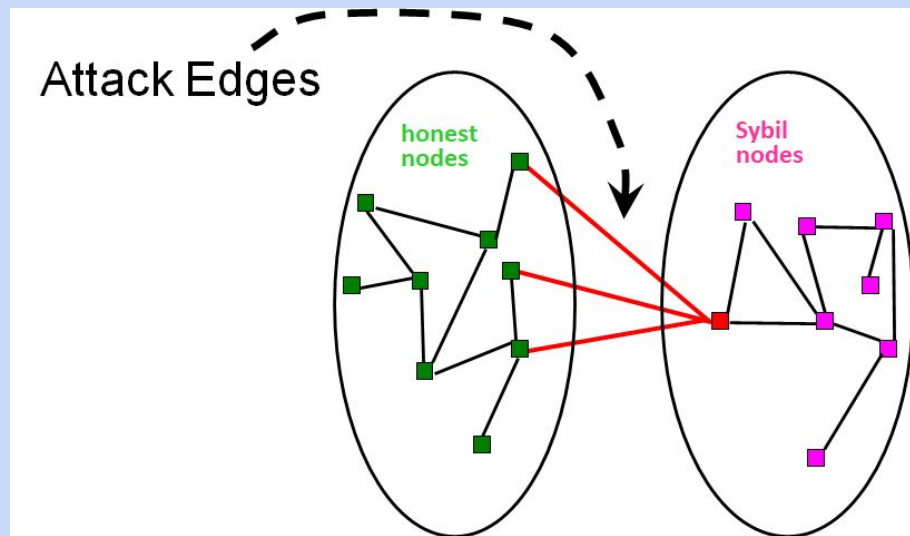
Our Social Network Definition



- ▶ Undirected graph
- ▶ Nodes = identities
- ▶ Edges = **strong** trust
 - E.g., colleagues, relatives

Sybil Nodes and Attack Edges

- Edges to honest nodes are “human established”
- Attack edges are difficult for Sybil nodes to create
- Attack edges are **rare**
 - To subvert system an attacker must compromise many honest nodes



SybilGuard

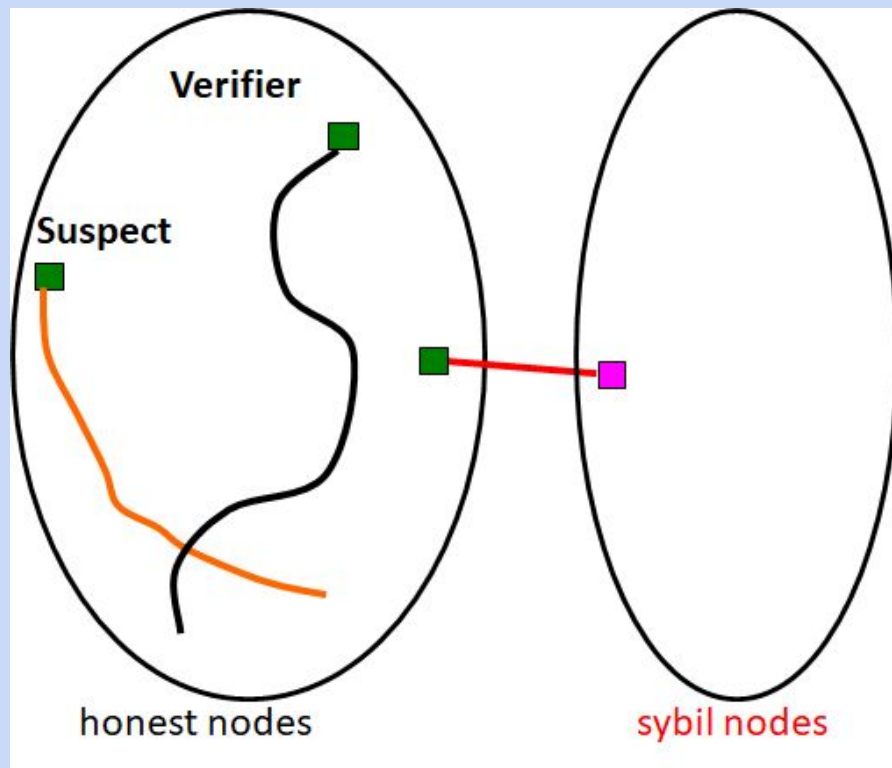
- A social network exists containing honest nodes and Sybil nodes
- Honest nodes provide a service to or receive a service from nodes that they “accept”
- Ideally, only honest nodes are accepted
- With high probability an honest nodes
 - Accepts most honest nodes
 - Is accepted by most honest nodes
 - Accepts at most a bounded number of Sybil nodes

Goal of Sybil Defense

- Goal:
 - Enable a **verifier node** to decide whether to accept another suspect node
 - Accept: Provide service to / receive service from
 - Idealized guarantee: An honest node accepts and only accepts other honest nodes
- SybilGuard:
 - Bounds the number of sybil nodes accepted
 - Guarantees are with high probability
 - Accepts and is accepted by most honest nodes
 - Approach: Acceptance based on random route intersection between verifier and suspect

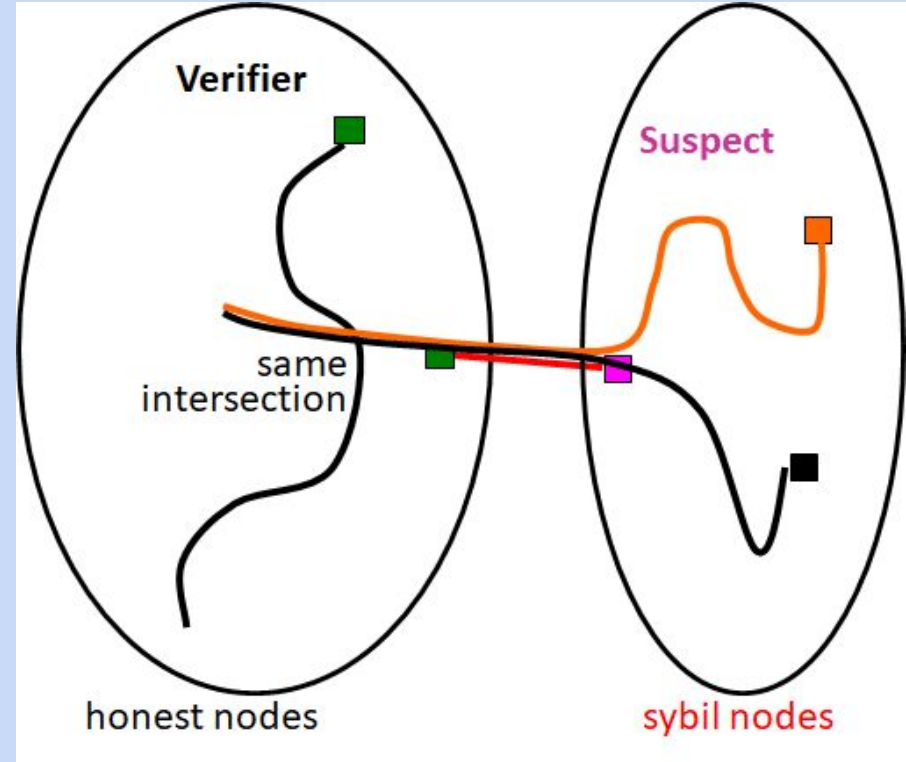
Random Route Intersection

- Random walk
 - Each node finds all the length w random routes that start at it
 - Honest node V accepts node S if most of V 's random routes intersect a random route of S
- With high probability
 - verifier's route stays within honest region
 - routes from two honest nodes intersect
- Verifier accepts a suspect if the two routes intersect



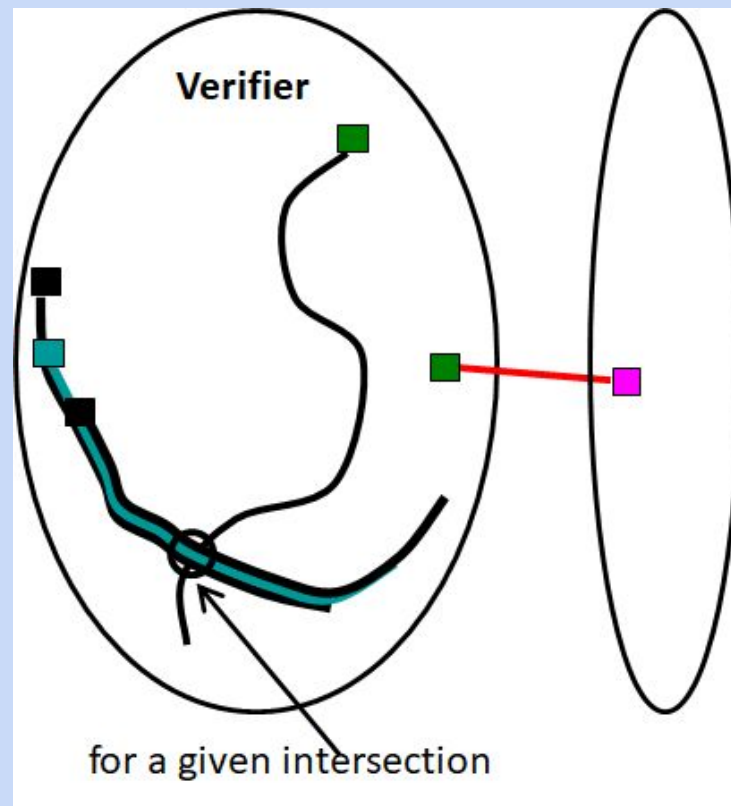
Random Route Intersection

- Each attack edge gives one intersection
- Intersection points are SybilGuard's equivalence sets



Random Route Intersection

- SybilGuard bounds the number of accepted sybil nodes within $g \cdot w$
 - g : Number of attack edges
 - w : Length of random routes



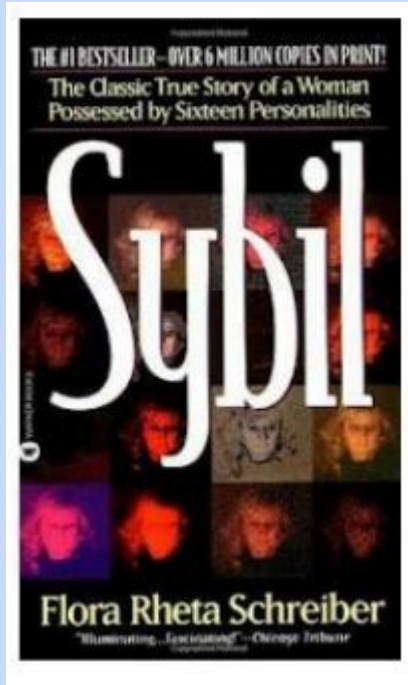
Uncovering Social Network Sybils in the Wild

Zhi Yang, Christo Wilson, Xiao Wang, Tingting Gao, Ben Y. Zhao, Yafei Dai

Peking University, UC Santa Barbara

IMC 2011

Sybil, fake account



Sybil, Noun

: a book of which content is a case study of a woman diagnosed with multiple personality disorder

“a fake account that attempts to create many friendships with honest users”

Target: Renren

- Renren: oldest and largest OSN in China



Previous detector on Renren

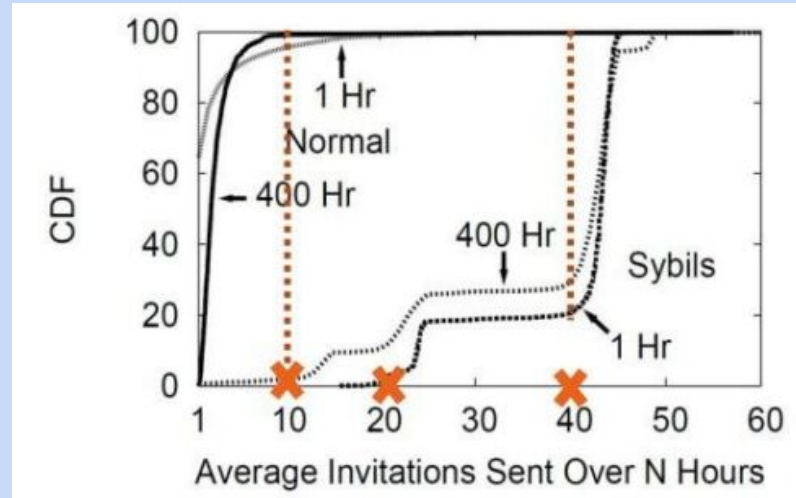
- Using orthogonal techniques to find sybil accounts
 - spamming & scanning content for suspect keywords and blacklisted URLs
 - crowdsourced account flagging
- Detect results
 - 560 sybils banned as of Aug 2010
- Limitations:
 - ad-hoc
 - require human effort
 - operate after posing spam content

Improved Detector

- Developed improved Sybil detector for Renren
 - Analyze ground-truth data on existing sybils
 - find behavioral attributes to identify sybil accounts
 - examine a wide range of attributes
 - find four potential identifiers

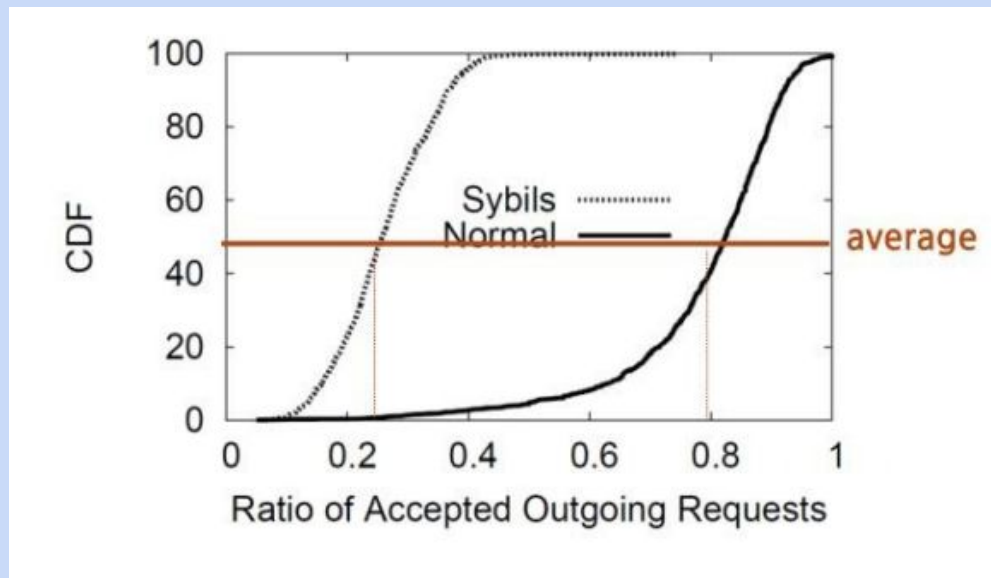
Four Reliable Sybil Indicators

- Friend request frequency (invitation frequency)
 - the number of friend requests a user has sent within a fixed time period



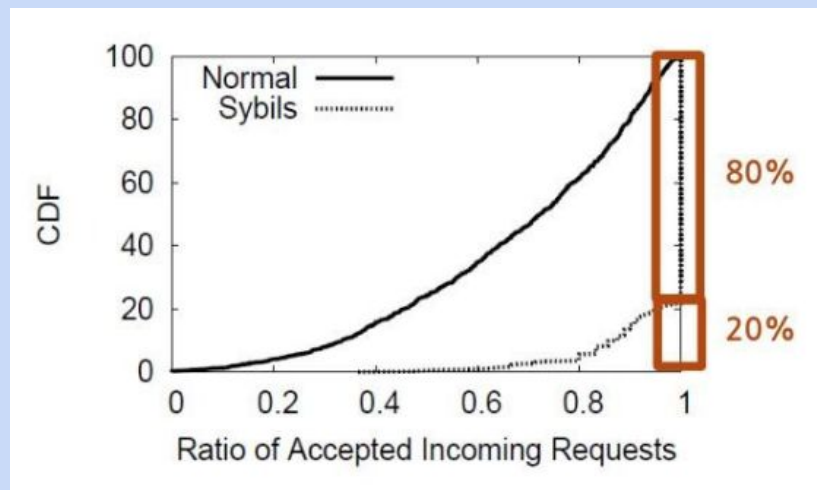
Four Reliable Sybil Indicators

- Outgoing friend request accepted
 - requests confirmed by the recipient



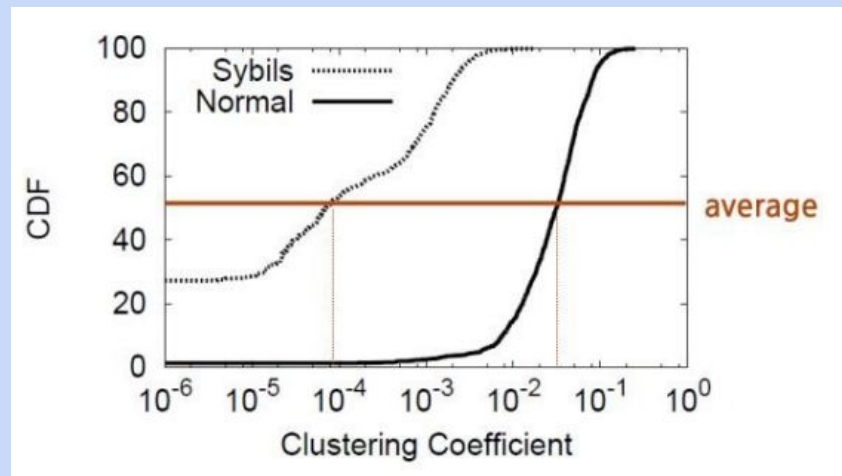
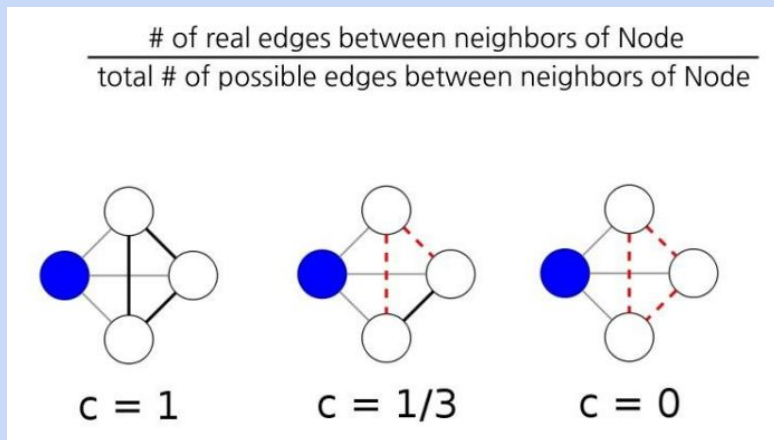
Four Reliable Sybil Indicators

- Incoming friend request accepted
 - The fraction of incoming friend requests accepted



Four Reliable Sybil Indicators

- Clustering coefficient
 - a graph metric that measures the mutual connectivity of a user's friends



Verify Sybil Detector

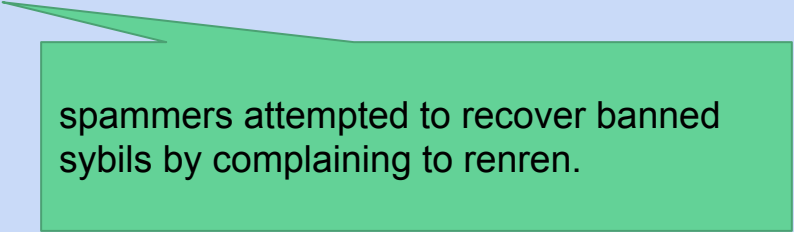
- Evaluate threshold and SVM detectors
 - dataset: 1000 normal user and 1000 sybils
 - similar accuracy for both

SVM		Threshold	
Sybil	Non-Sybil	Sybil	Non-Sybil
98.99%	99.34%	98.68%	99.5%

- deployed threshold, less CPU intensive, real-time
- adaptive feedback scheme is used to dynamically tune threshold parameters

Detection Results

- Detect 100K sybils in the first six months (aug 2010 - feb 2011)
 - vast majority (67%) are spammers
- Low false positive rate
 - use customer complaint rate as signal
 - complaints evaluated by humans
 - 25 **real** complaints per 3000 bans (<1%)



spammers attempted to recover banned sybils by complaining to renren.