The Latent Community Model for Detecting Sybil Attacks

Zhuhua Cai, Christopher Jermaine
Rice University
{zc7, cmj4} @ rice.edu

Outline

I Sybil attack

II Current solutions

III Our approach

Sybil Attack

An attacker creates multiple fake identities (Sybils) to gain influence in the open system

… …

Examples

- Recommendation system, i.e., Drugstore …
- Email system, (spam email)
- Web spam
- Distributed Hash Table (DHT)
- Communication System (Tor)

… …

YouTube | Rice University | Search | Browse | Upload | Create Account | Sign In

Rice University: A Campus Tour

NigelCoxworth | 41 videos | Subscribe

0:07 / 360p

Like

15,847

29 likes, 2 dislikes

Uploaded by Nige

An architectu ng Weekend 2006.

Song by: Hem

Buy this song:
iTunes
AmazonMP3

Influence

Show more

Uploader Comments (NigelCoxworth)

Very nice video. Well put together - very professional. If you are professional that is not intended as insult I am merely ignorant. At homecoming 2006 I was a senior. It was a particularly poignant year for me full of hope, wonder, fear, regret, and sorrow. It was a struggle to hold back my tears watching my Alma mater portrayed so delicately. Thanks. - Paul Hanszen College 2007

paulgrutherford 3 years ago

Suggestions

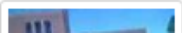
Calling All Engineers
by geecochallenge
822 views Promoted Video

Rice University Virtual Tour
by DOESolarDecathlon
6,639 views

Rice University Tips and Tricks -Are Grades Rea...
by OverThere1142
2,483 views

Rice University - Orientation Week (O-Week) Out...
by RiceUniversityCIC
1,919 views

Rice Student Association
by RiceUniversity
953 views



Rice University Bike Ride

Outline

I Sybil attack

II Current solutions

III Our approach

Current Solutions

- Prevention
- Detection

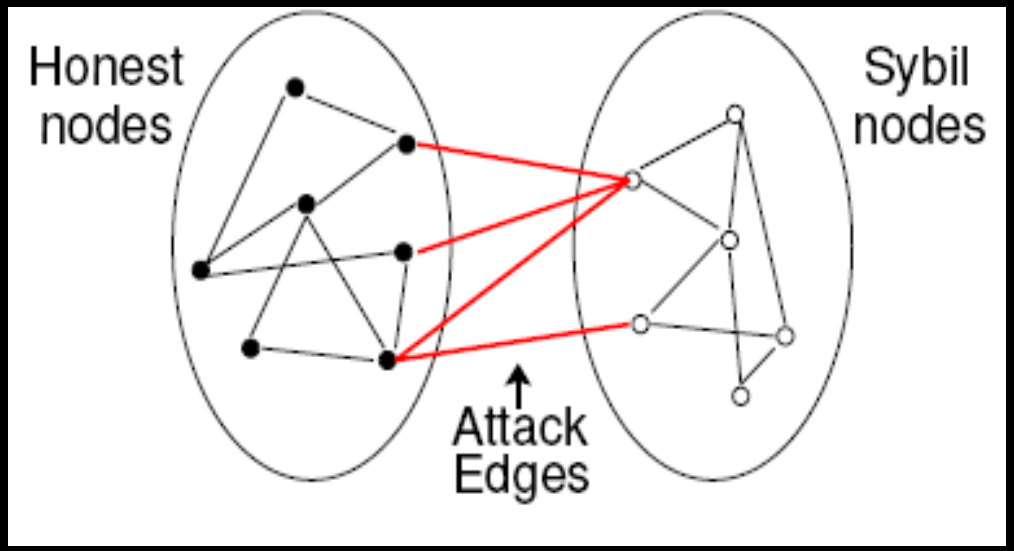
Solutions / Prevention

- Challenge/Response Mechanisms
  - Computational Puzzle or CAPTCHA
- Credentials
  - E.g., social security number, driver license, banking account

Solutions /Detection

- Trust/Reputation
  - Amazon's seller rating system
  - be subjected to Whitewashing attack
- IP/IP-Clustered
  - be subjected to Botnet
- Machine Learning
  - Features like invitation frequency, requests, etc
- PageRank and HITS
  - Trusted pages

Using Social networks



- Bottleneck Cut
- Fast Mixing Property

- SybilGuard [SIGCOMM'06], SybilLimit [Oakland'08], SybilInfer [NDSS'09], SumUp [NSDI'09], DSybil [Oakland '09], GateKeeper [Infocom' 10]

Outline

I Sybil attack

II Current solutions

III Our approach

Our Approach

Rather than assuming the forms of attacks, we instead learn a statistical generative model for the underlying network, called the *latent community (LC) model.*

What is Generative Model ?

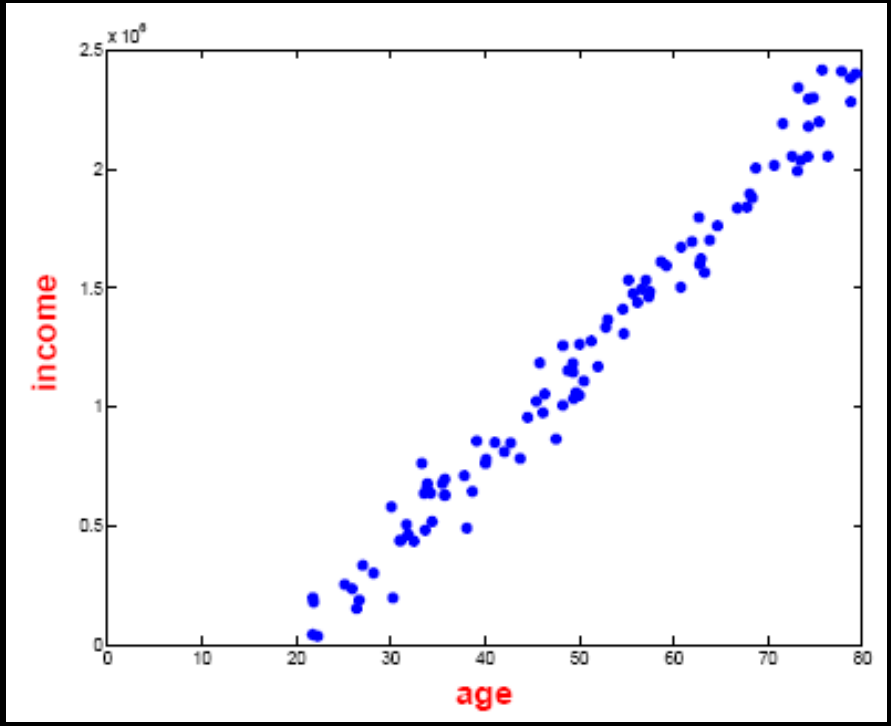
A **generative model** is a model that describes the sequence of distributions that generates our observable data.

Example

- Regression

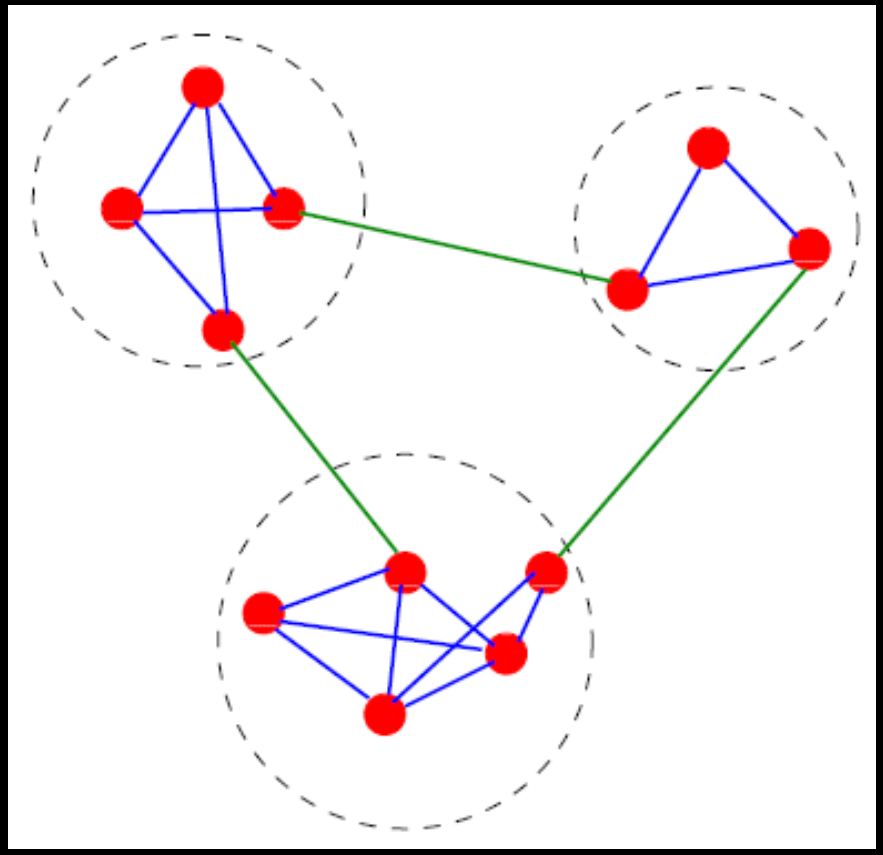
$$y \sim Norm(a * x + b, \sigma^2)$$

  - y: income, x : age
  - Given an x, you can estimate: y and its probability



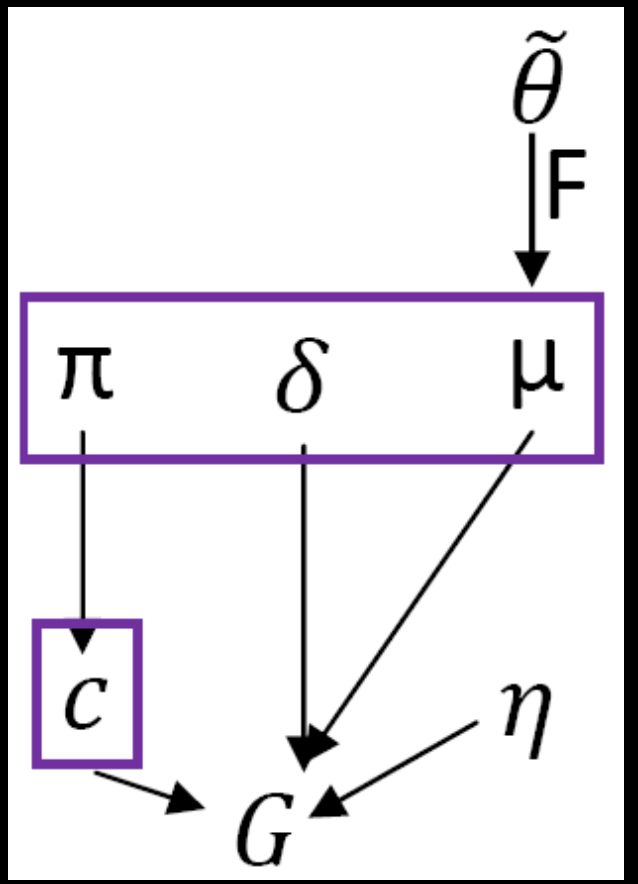
What is LC Model?

- LC is a Statistical generative model
  - Node
  - Edge
  - Community
  - Latent positions

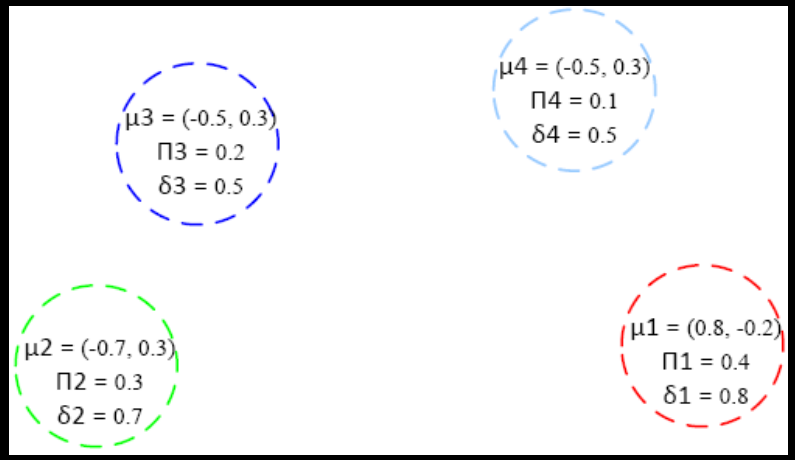


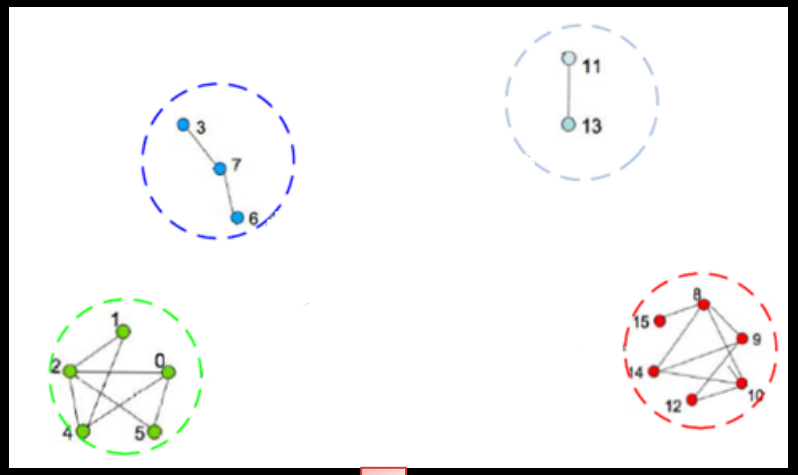
LC model

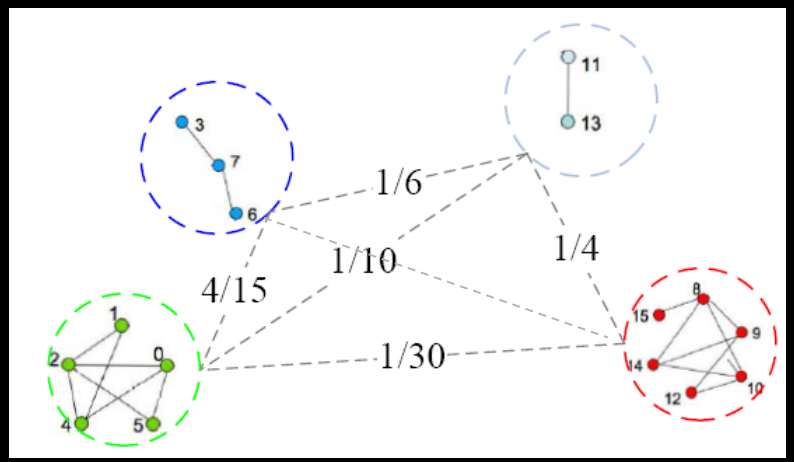
- Bayesian network of LC model
  - π: the fraction of nodes in each community.
  - δ: the internal edge density
  - μ: the positions in Euclidean space.
  - c: the membership of nodes
  - η: a scaling factor

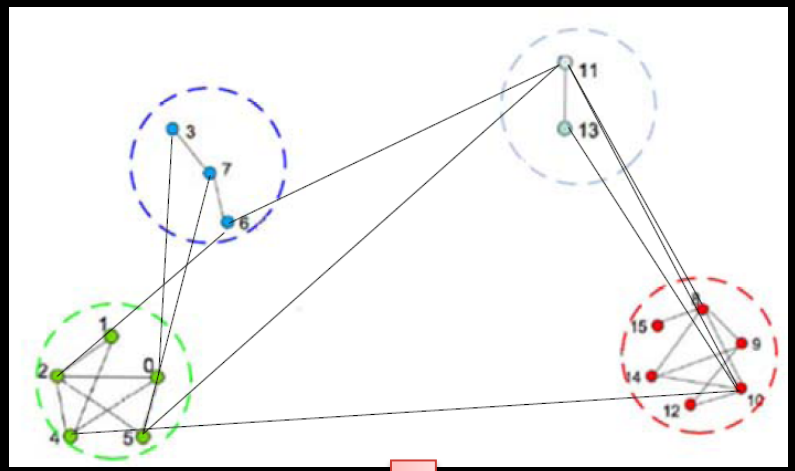


An Example

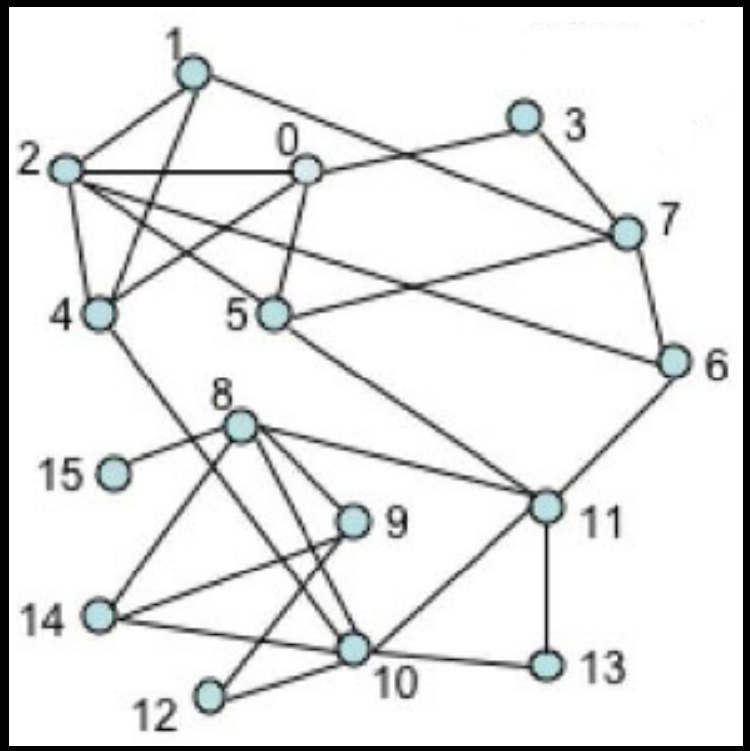

μ3 = (-0.5, 0.3)
Π3 = 0.2
δ3 = 0.5
μ4 = (-0.5, 0.3)
Π4 = 0.1
δ4 = 0.5
μ2 = (-0.7, 0.3)
Π2 = 0.3
δ2 = 0.7
μ1 = (0.8, -0.2)
Π1 = 0.4
δ1 = 0.8

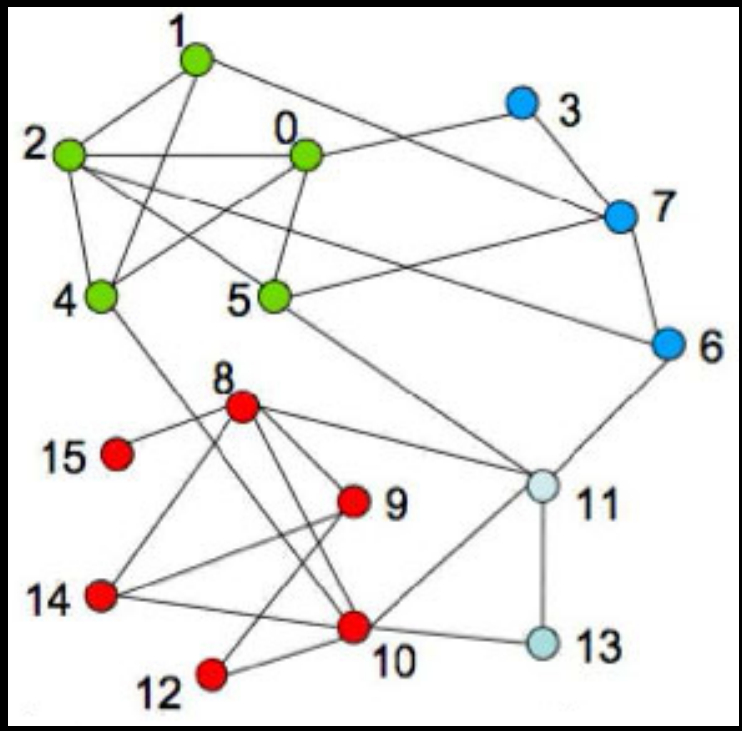

11
13
3
7
6
1
2
0
4
5
8
15
9
14
12
10


11
13
3
7
6
1/6
1/10
1/4
4/15
1/30
1
2
0
4
5
8
15
9
14
12
10

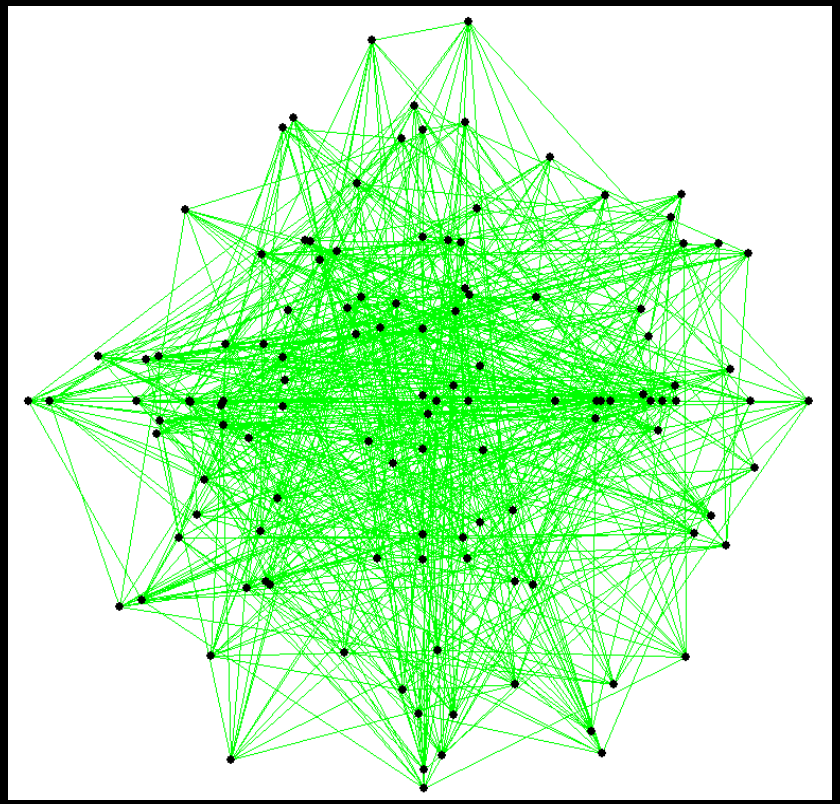

11
13
3
7
6
1
2
0
4
5
8
15
9
14
12
10

Learning Process





Second Example



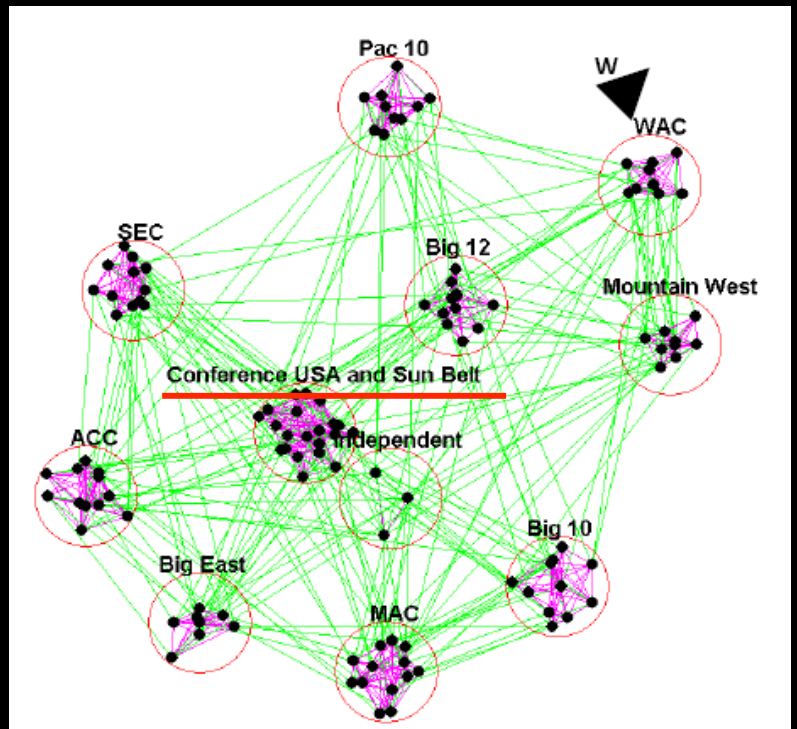
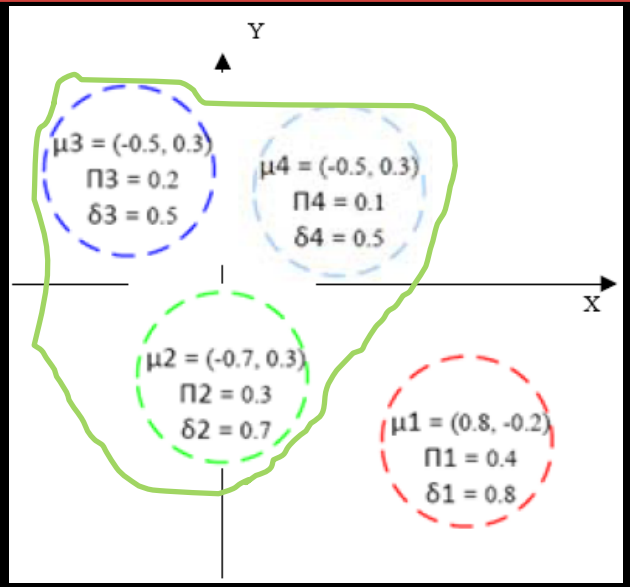
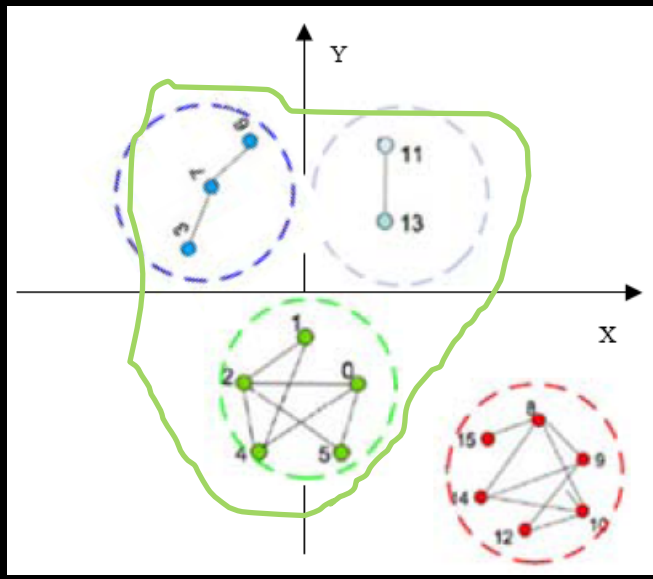


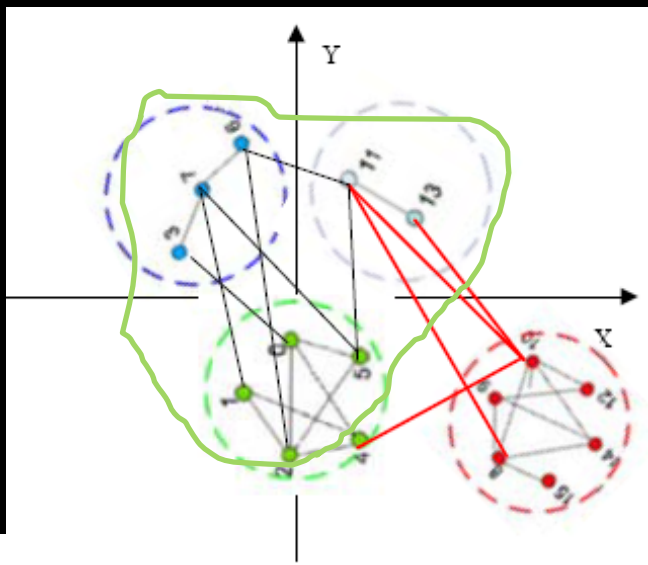
Figure 1: Learned model for the 2010 college football data set.

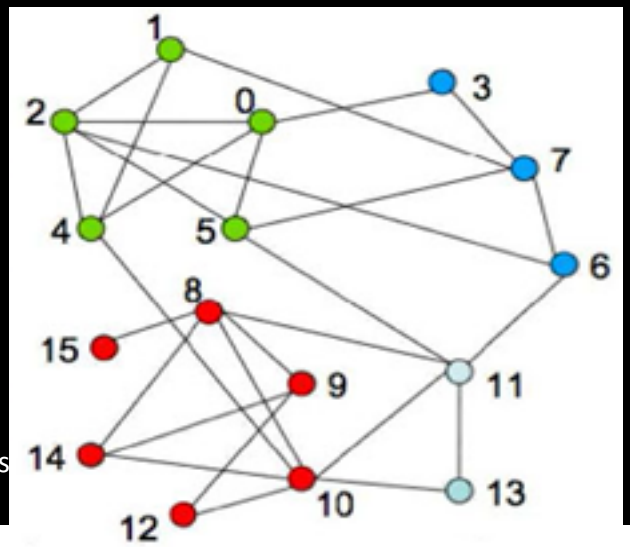
2010 American college football schedule

LC-based Sybil Detector (1/2)



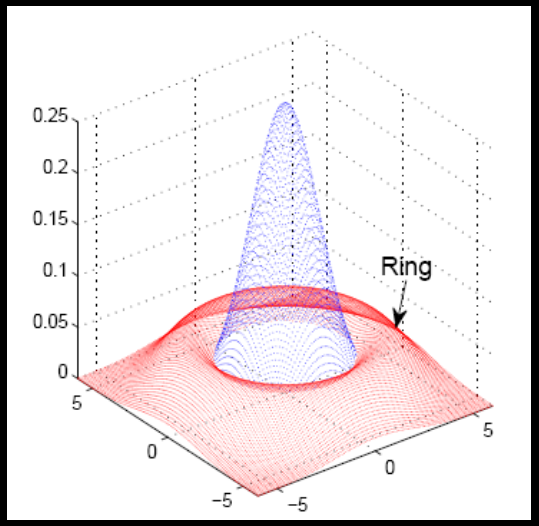


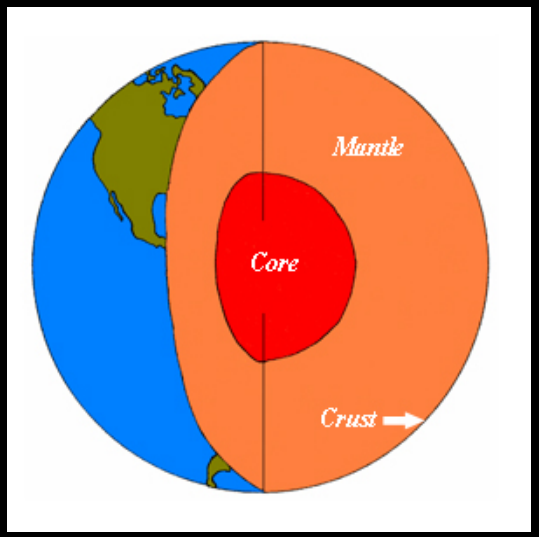




LC-based Sybil detector (2/2)

- Two kinds of communities
- Assumptions:
  - Seeds
  - Nodes in the same community have same properties.





Bayesian Inference Engine for learning algorithm

- Learning Algorithm

$$P(\Theta|G) = \frac{P(G|\Theta)P(\Theta)}{P(G)}$$

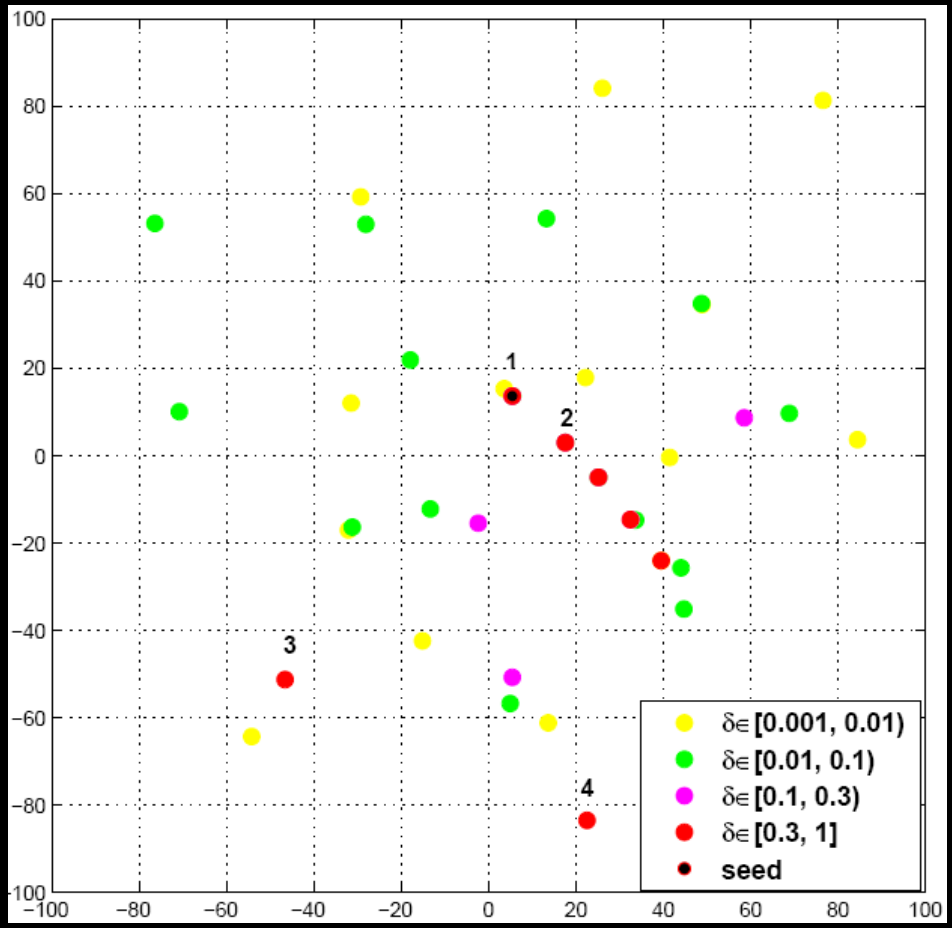
- Gibbs Sampling
  1. Choose initial values.
  2. Iterate over each parameter, and sample values.
  3. Aggregate the distribution of target parameter.

LC Model on Digg (1/4)

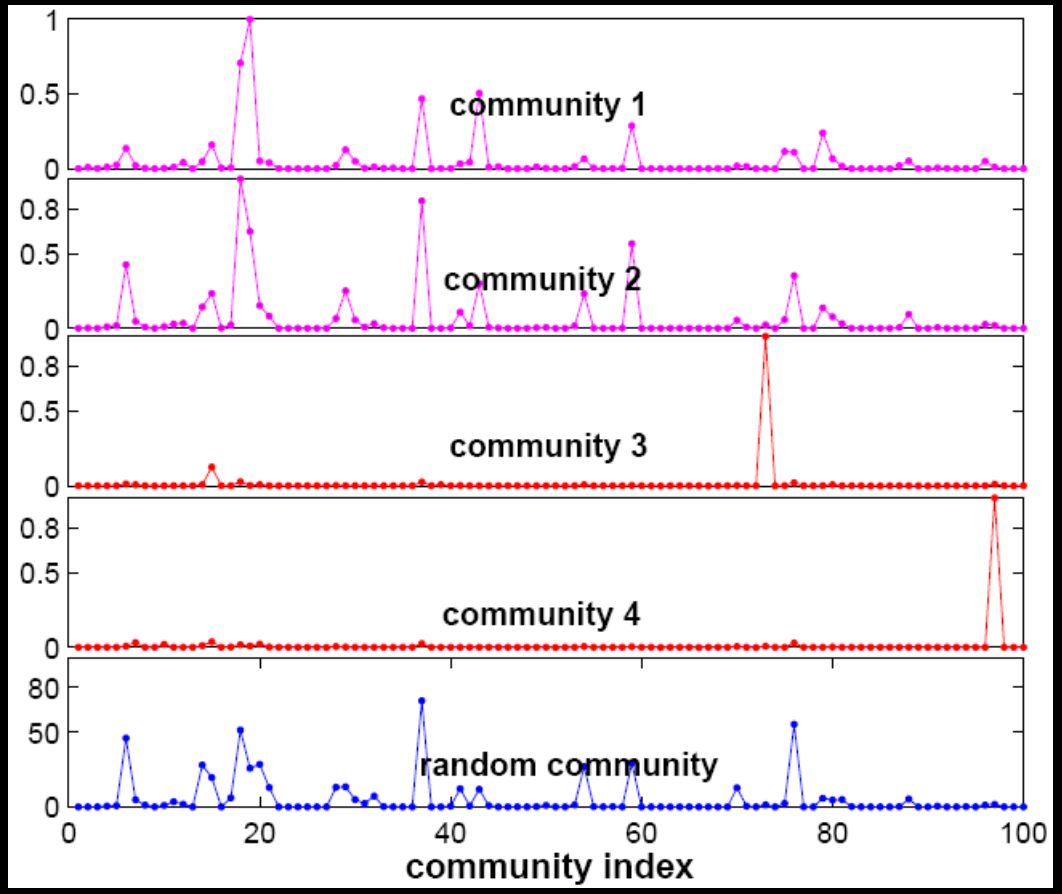
- Digg
  - Following or being followed by others
  - Digg or Bury others
  - Motivation for Sybils
- Dataset
  - 594, 426 nodes
  - 5, 066, 988 edges

LC Model on Digg (2/4)

- Configurations
  - 100 clusters
  - "Kevin Rose" as a seed
  - 200 cycles
- Sybil communities
  - Community 3 and 4
  - δ: 0.40 and 0.55
  - n: 311 and 299

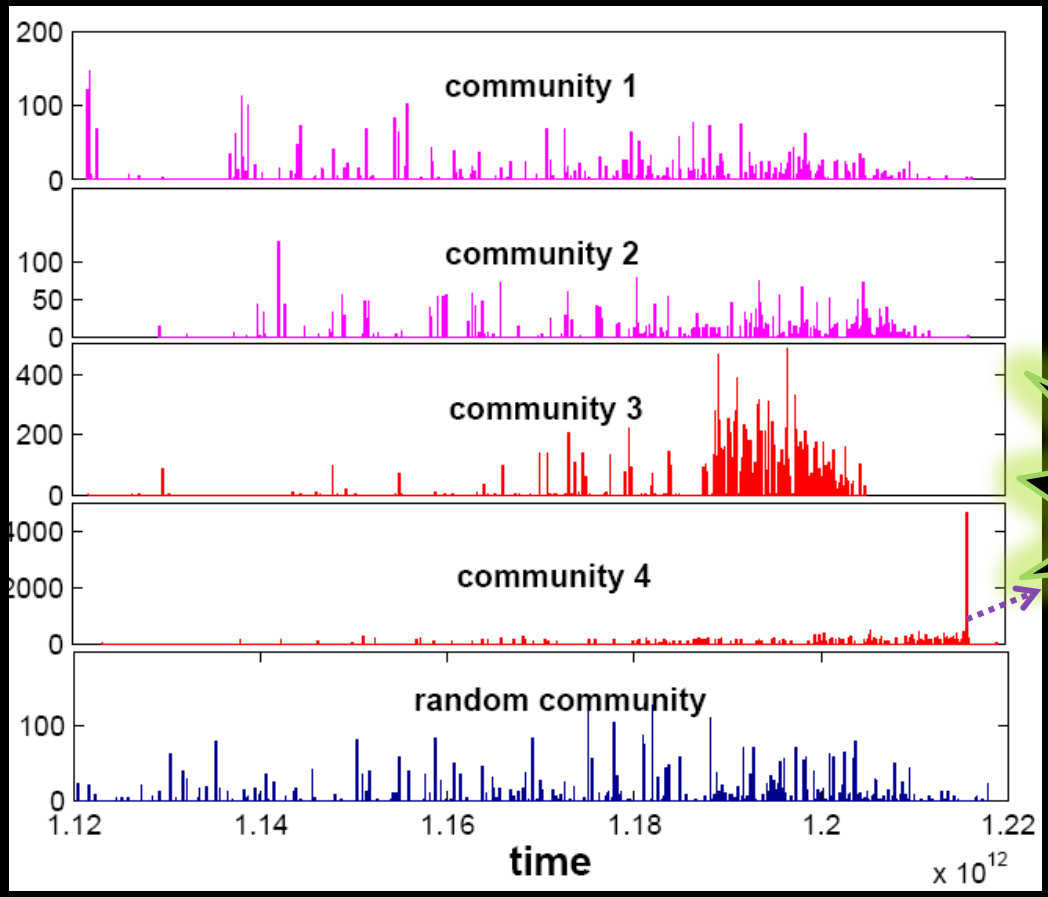


LC Model on Digg (3/4)



The relative edge density among Digg communities

LC Model on Digg (4/4)



3 days a**nd** **then stops!**

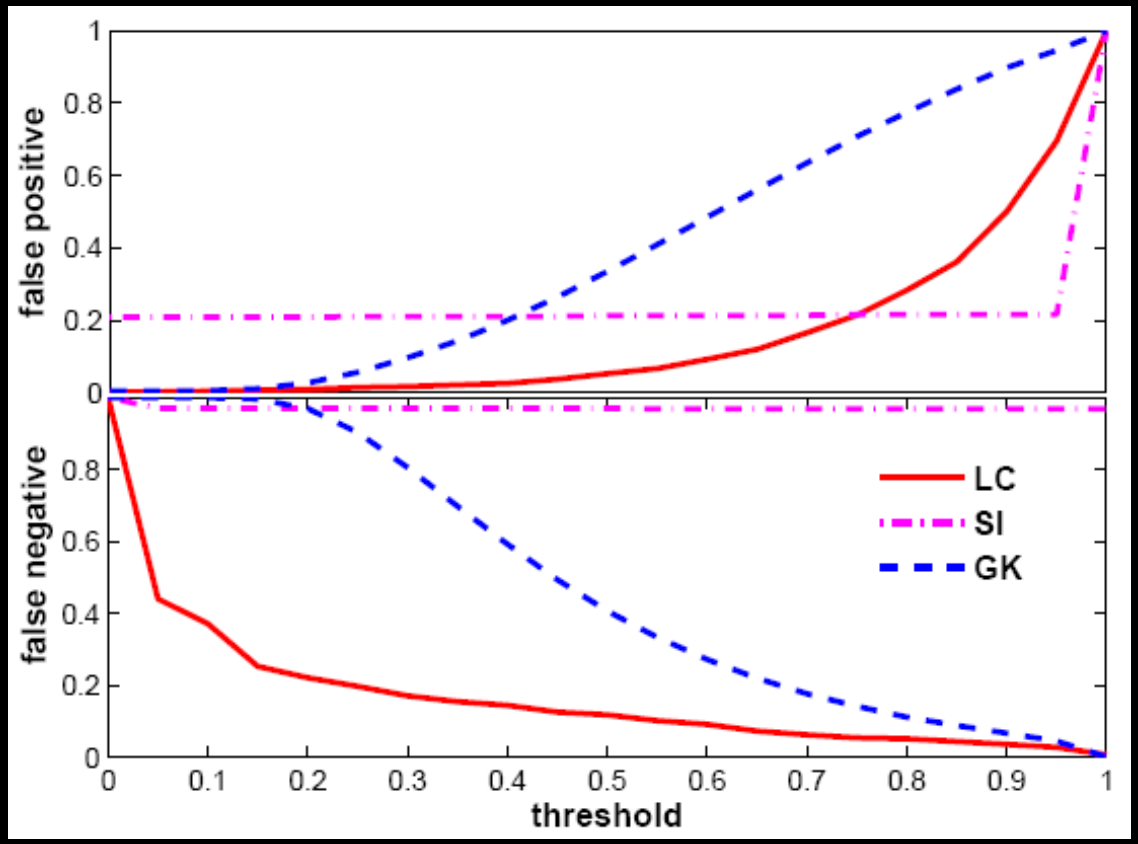
The creation time of edges in Digg communities

LC-based Sybil Detector (1/2)

- Compared Algorithm
  - SybilInfer and GateKeeper
- Simulate Sybil attacks
  - Attackers, victims, seeds, attack topologies
- Datasets

Dataset	Node	Edge	Directed
Irvine Community	1899	13, 820	True
Wikipedia Vote	7115	100, 762	True
Gnutella	8717	31, 525	False
Email-Enron	36, 692	367, 662	False

LC-based Automatic Sybil Detector (2/2)



General result for comparison

Discussion

- Algorithm Complexity
  - O(k*n) (when the number of communities is k).
  - 200 cycles lead to result
- LC model in other literatures
  - Content distribution in clusters
  - Geographical applications
  - … …

Conclusion

- LC model shows good performance on the Sybil problem
- Weakness
  - Tree-topology attack or Sparse attack.
  - It is not a distributed algorithm.
  - It is not used in applications without social networks.

Q & A

Thank You !