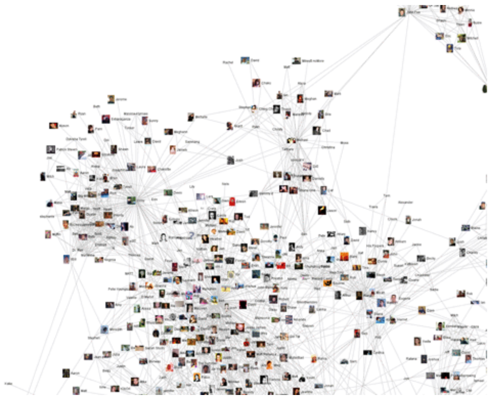
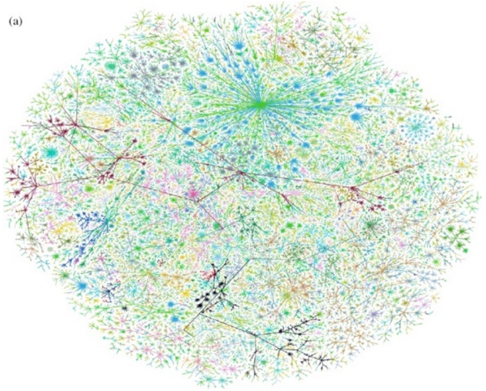
Poisoning Attacks to Graph-based Methods

Neil Gong

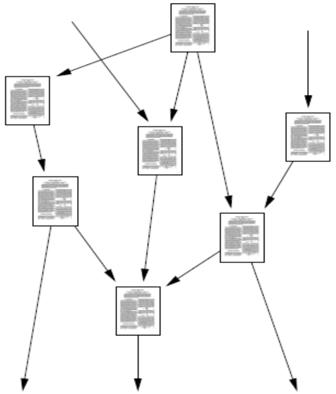
Graphs are Ubiquitous



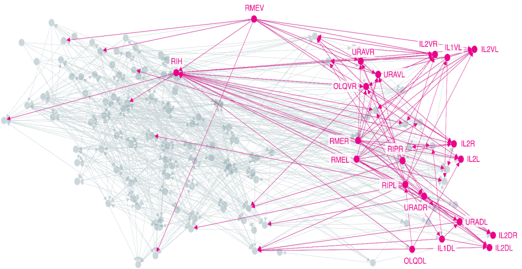
Social networks



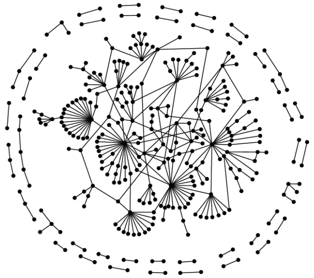
Internet



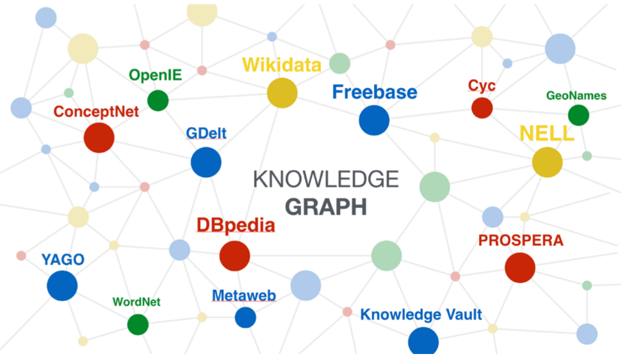
Citation network



Neuron Networks

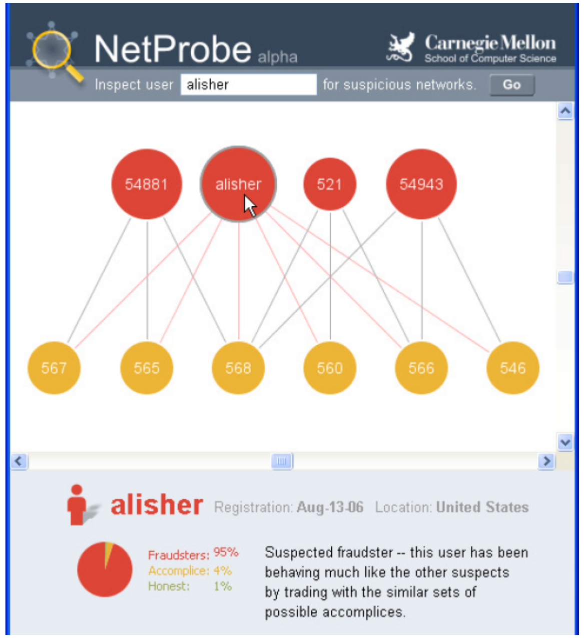


Biomedical networks

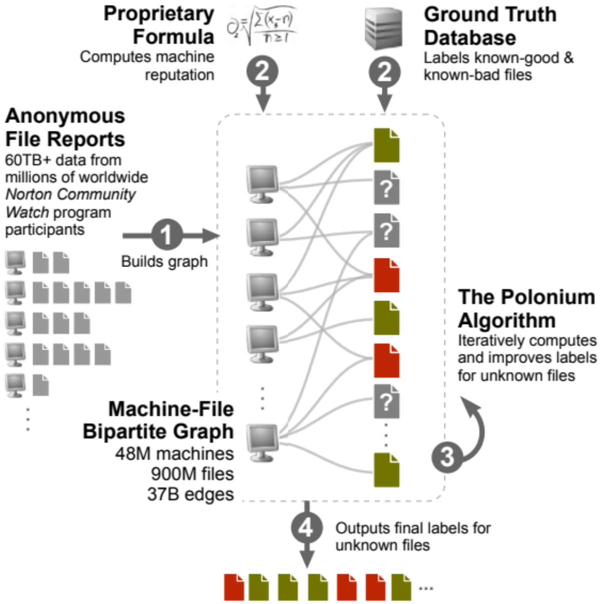


Graph-based Security Applications

Malicious user detection
in social networks



Fraud detection in
online auction network



Malware detection in
machine-file graph

Node Classification

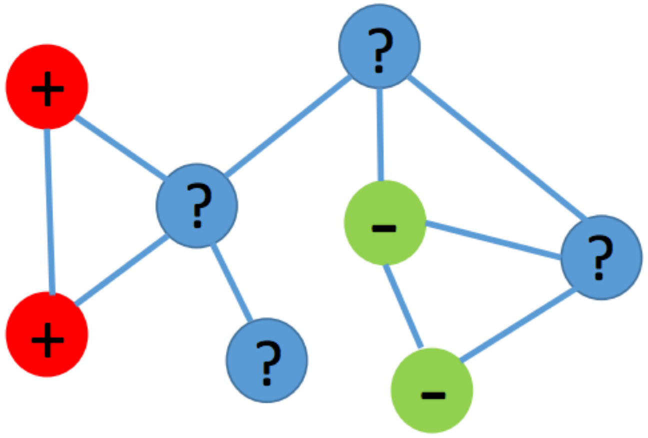
- Conventional methods
  - Random Walk (RW)
  - Loopy Belief Propagation (LBP)
  - Linearized Loopy Belief Propagation (LinLBP)
  - …

2011 ACM Turing Award

- Graph Neural Network
  - Graph Convolutional Network (GCN)
  - Graph Attention Network (GAT)
  - GraphSAGE
  - …

Node Classification

- Input
  - Undirected (or directed) graph
  - Node/edge features (optional)
  - Training set
    - Labeled positive nodes (+)
    - Labeled negative nodes (-)

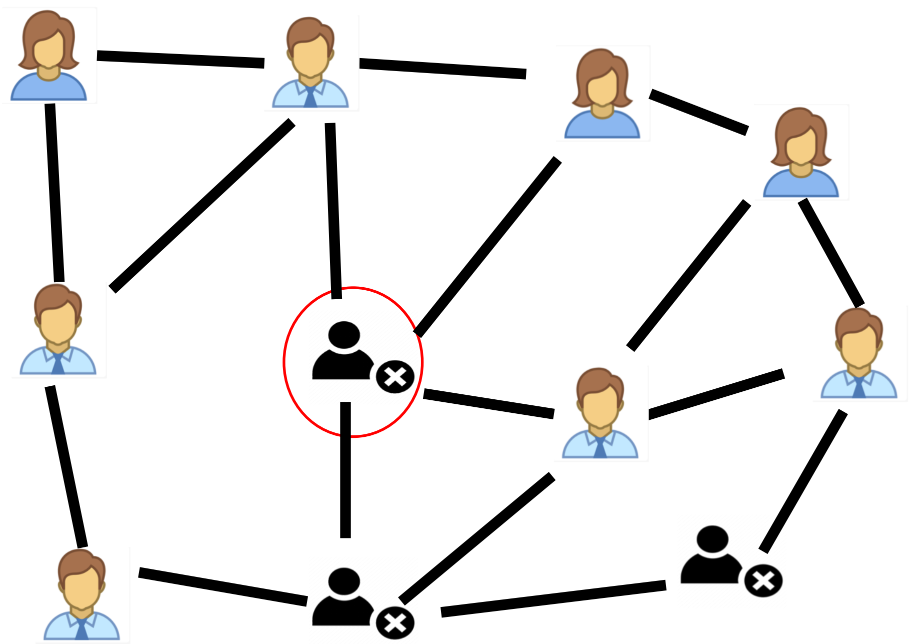


- Output
  - Estimate labels of unlabeled nodes (?) *simultaneously*

Attacks to Graph-based Classification

Attacks to Graph-based Classification

- Threat Model
  - Attacker's knowledge
  - Attacker's capability
  - Attacker's goal
    - Attacker's target nodes (malicious) are misclassified as normal users



Attacker’s Knowledge

- Imagine you are a malicious user in social network (e.g., Facebook)
  - Facebook leverages graph-based classification method to detect malicious users

- Whether knowing Complete Graph

- Whether knowing Training Dataset

- Whether knowing Model Parameters

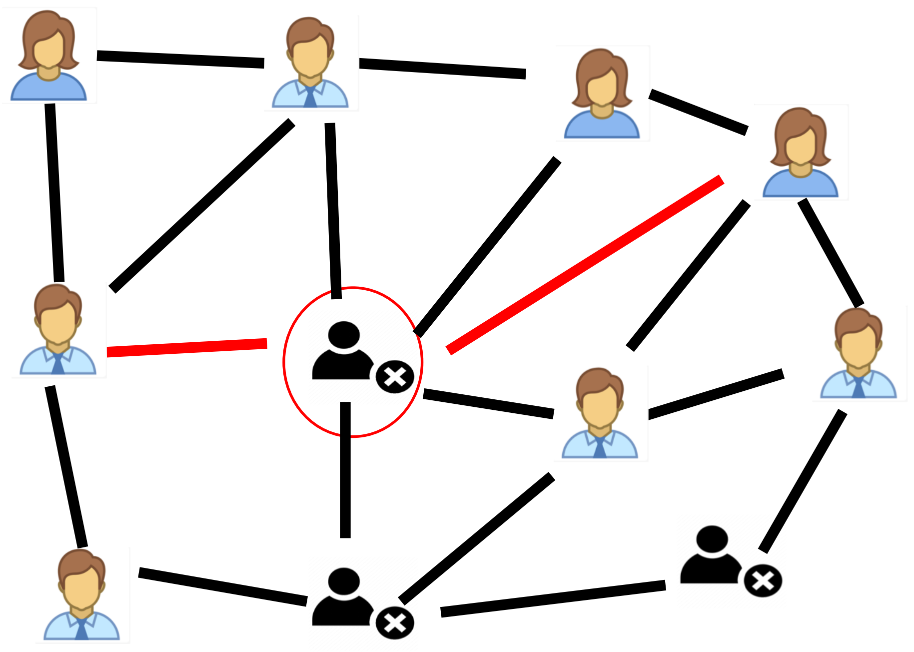
Attacker’s Capability

- Way 1: Modify the target node’s features
  - A malicious user can modify his profile so as to resemble benign user’s
- Way 2: Modify the target node’s local structure (add/delete edges)
  - A malicious user can buy followers or unfollow users
- Way 3:Modify both target node’s features and local structure
  - A malicious user can modify both his profile and buy followers/unfollow users

Attack Strategy

- Random attack

Random add/remove edges between target node and other nodes.

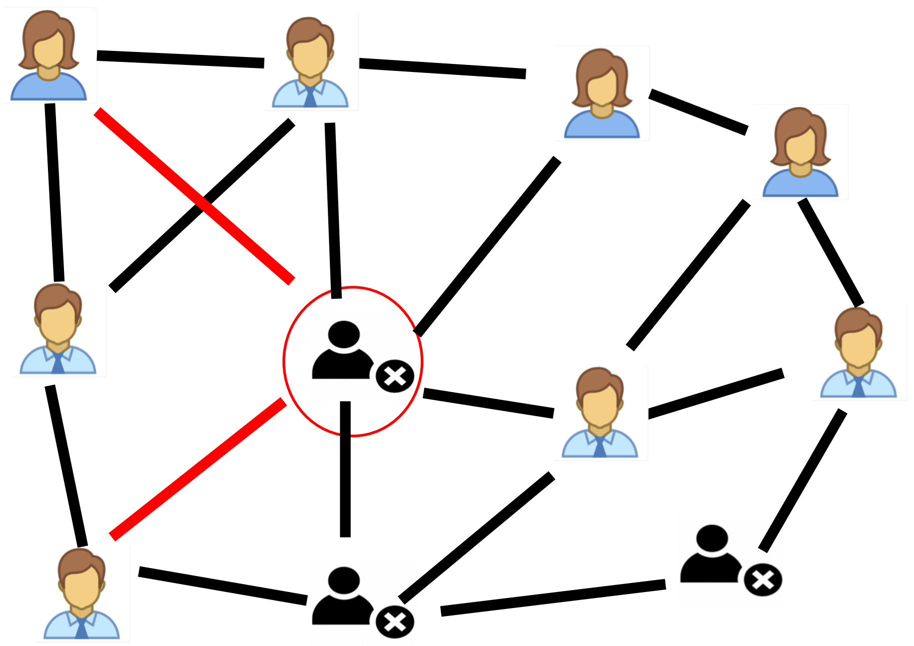


Attack Strategy

- Delete-Add attack

  First delete edges between target node and its connected positive nodes

  Then add edges between target node and random selected negative nodes



Formulating Attacks as Optimization Problems

- Attacker's knowledge: Compete graph, training set, model parameters
- Attacker's capability: modify target nodes' local structure
- Attacker's goal: misclassify attacker's target nodes (FNR=1)

$$\min_{\mathbf{B}} \sum_{u,v \in V, u<v} B_{uv} C_{uv}, \longrightarrow$$ Minimize total cost on all pairs of nodes

$$\text{s.t.} \quad FNR = 1, \longrightarrow$$ Misclassify attacker's target nodes

$$B_{uv} \in \{0,1\}, \ \text{for } u, v \in V, \longrightarrow$$ B_{uv} binary variable

$$\sum_{v} B_{uv} \leq K, \ \text{for } u \in V, \longrightarrow$$ Maximum number of modified edges

Adversarial matrix B: $B_{uv} = 1$ means modifying the connection status between *u* and *v*
Cost matrix C: C_{uv} is the cost of modifying the connection status between *u* and *v*

Optimization-based Attack vs. Heuristic Attacks

Dataset	No attack	Random attack	Del-Add attack	Our attack
	FNR	FNR	FNR	FNR
Facebook	0	0.02	0.43	0.94
Enron	0	0.03	0.76	1.00
Epinions	0	0.02	0.63	0.99
Twitter	0	0.02	0.43	0.88

Attacks to Different Methods

Method	**GCN**	**LINE**	**RW**	**LBP**	**JWP**	**LinLBP**	**Time**
No attack	0.05	0.01	0.03	0.01	0	0	0 sec
Nettack	**0.64**	0.58	0.33	0.28	0.13	0.22	9 hrs
Our attack	0.54	**0.85**	**0.92**	**0.92**	**0.93**	**0.94**	**10 secs**