AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

19CSE337 Social Networking and Security

Lecture 25

Topics to Discuss

- Mitigating Sybil Attacks.
- Mitigating Social Spam.

Mitigating Sybil Attacks

- The popularity of modern social networks has rendered social media platforms vulnerable to malicious activities.
- One such activity is a Sybil attack, in which a single entity emulates the behaviors of multiple users and attempts to create problems for other users and a network itself.
- YouTube, Facebook, and BitTorrent have become vulnerable to sybil attacks.

Mitigating Sybil Attacks

- Two methods to mitigate sybil attacks.
  - **Sybil detection**: Sybil detection schemes use social graph structure to identify a given user is sybil or non-sybil.
  - **Sybil resistance**: It won't classify sybil or non-sybil, but use application specific knowledge to mitigate the influence of sybils in the network.

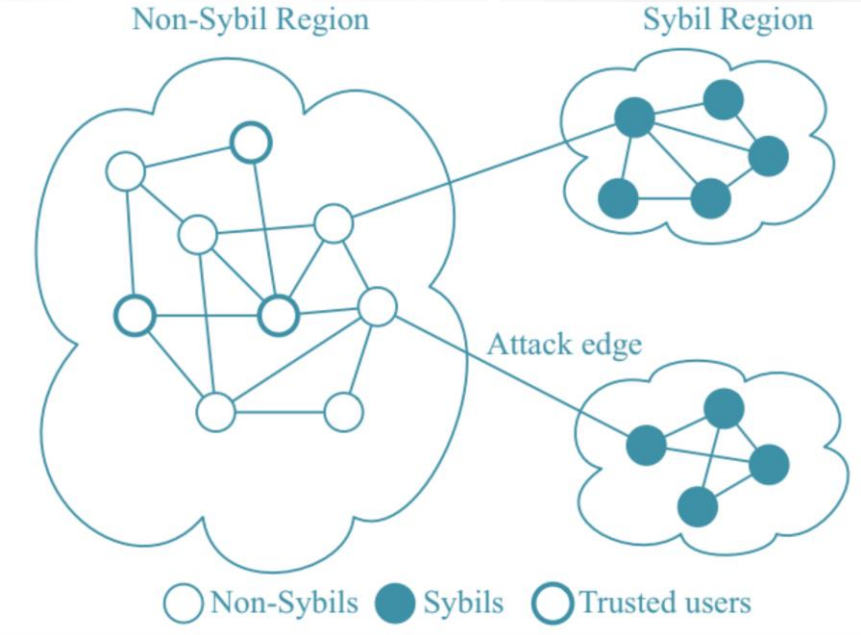
Sybil Detection

- The network is represented as a social graph G=(V,E).
- V represents users and E represents a connection.
- If a sybil node successfully able to connect with a non-sybil node then the edge is called attack edge.
- Attackers can launch sybil attacks by creating many sybil identities and creating attack edges with non-sybil users.

Common assumptions of sybil detection

- Attackers can create large number of sybil identities but they lack trust relationship and thereby limited number of attack edges.
- A limited number of attack edges causes sparse cut between sybil and non-sybil regions. Non-sybil regions are well connected.
- The defense mechanism knows at least one non-sybil.

Sybil Detection



Sybil Detection

- Since the non-Sybil region of the network is densely connected (assumption 1), and the Sybil region of the network is attached by a limited number of links (assumption 2), existing detection schemes look for resulting topological features to partition the network into Sybil and non-Sybil regions.
- They then look for the partition that contains the known non-Sybil identity (assumption 3) to decide which is the non-Sybil region.

Other Techniques

- Apply a random walk from non-sybil node to all other nodes.
- Apply community detection algorithms, will end up with sybil and non-sybil regions.
- Rank nodes based on trust computing.
- Apply betweenness centrality. High score edge can be an attack edge.

Sybil Resistance

- Useful to reduce the impact of sybil nodes in the network.
- Sybil resistance scheme uses additional parameters like user's interactions, transactions, votes etc.
- Based on the underlying structure and properties, sybil resistance scheme determine whether the action performed by a user should be allowed or not.

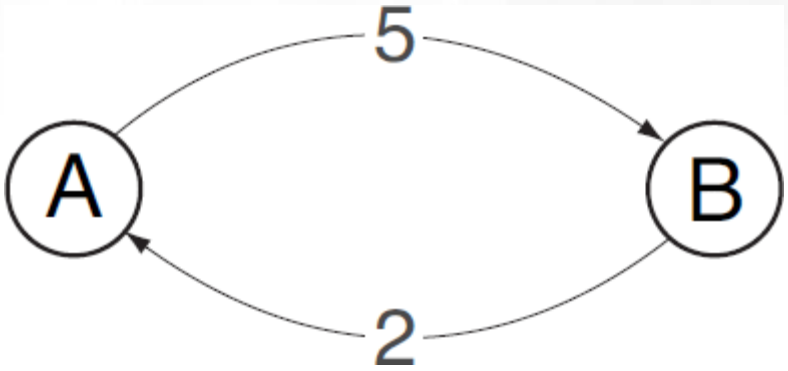
Sybil Resistance

- The sybil resistance scheme uses a credit network mechanism in which a pair of nodes assigns a credit value indicating their trust value.
- This credit score is used as the reference to perform certain action (such as sending message, receiving message etc) in the network.
- If the credit score is lesser than actual score required to perform a transaction, then the transaction can’t be performed or not allowed.

Credit Network

- A credit network can be formed from existing social network where each undirected edge is represented as two directed edges.
- Further, each directed edge, (a,b), is assigned an initial credit allocation C_{ab} by the destination node b.
- When a new social link is created, the requesting node should be required to grant the accepting node some initial credit but not vice-versa, to prevent an attacker from obtaining credit by initiating social links.

Credit Network



Simplified credit network between two nodes A and B, with credit available c_{ab} and c_{ba} shown. In this example, A has 5 credits available from B, and B has 2 credits available from A.

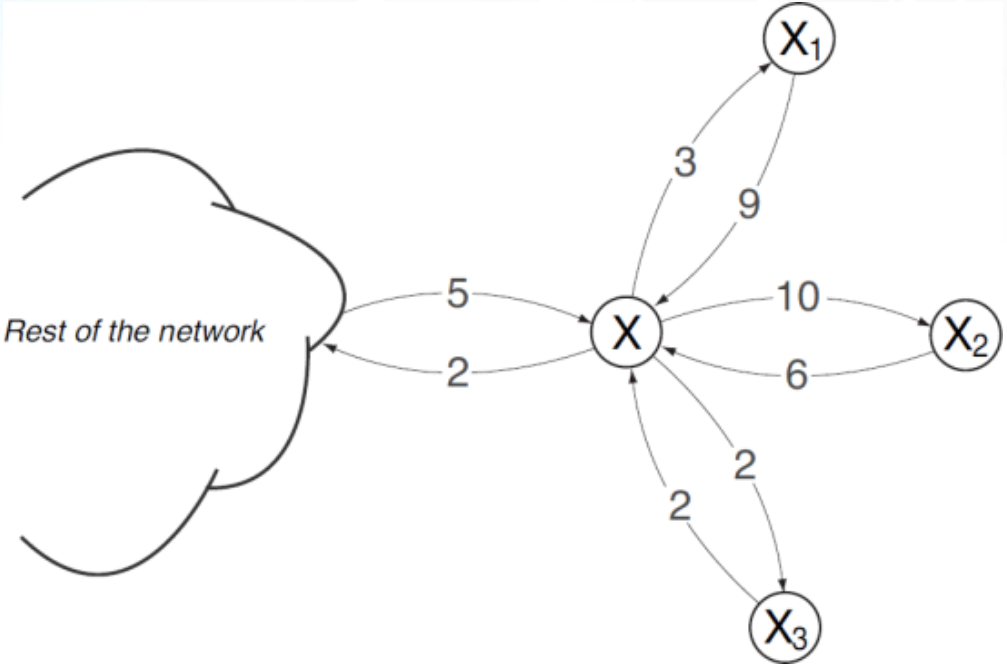
Credit Network

- An attacker can mount a Sybil attack by creating many different identities in the social network, each corresponding to a different node in the credit network.
- However, as per assumptions about credit assignment to links, having many user accounts does not by itself allow the attacker to obtain additional available credit with other users.

Credit Network

- The total amount of credit available to a single user is the sum of the credit available on her links to other users.
- An attacker with an arbitrary number of Sybil identities has exactly the same available credit as the attacker with just one identity.
- In this case, the relevant set of edges is the cut between the subgraph consisting of the attacker's Sybil identities and the rest of the network.
- Any credit available on edges between the attacker's Sybil identities does not matter, because it does not enable additional “purchases” from legitimate nodes.

Credit Network



Credit networks leading to Sybil tolerance. User X can create any number of identities (X1, X2, X3) and arbitrarily assign the credit available between them. However, does not enable any additional available credit with nodes in the rest of the network.

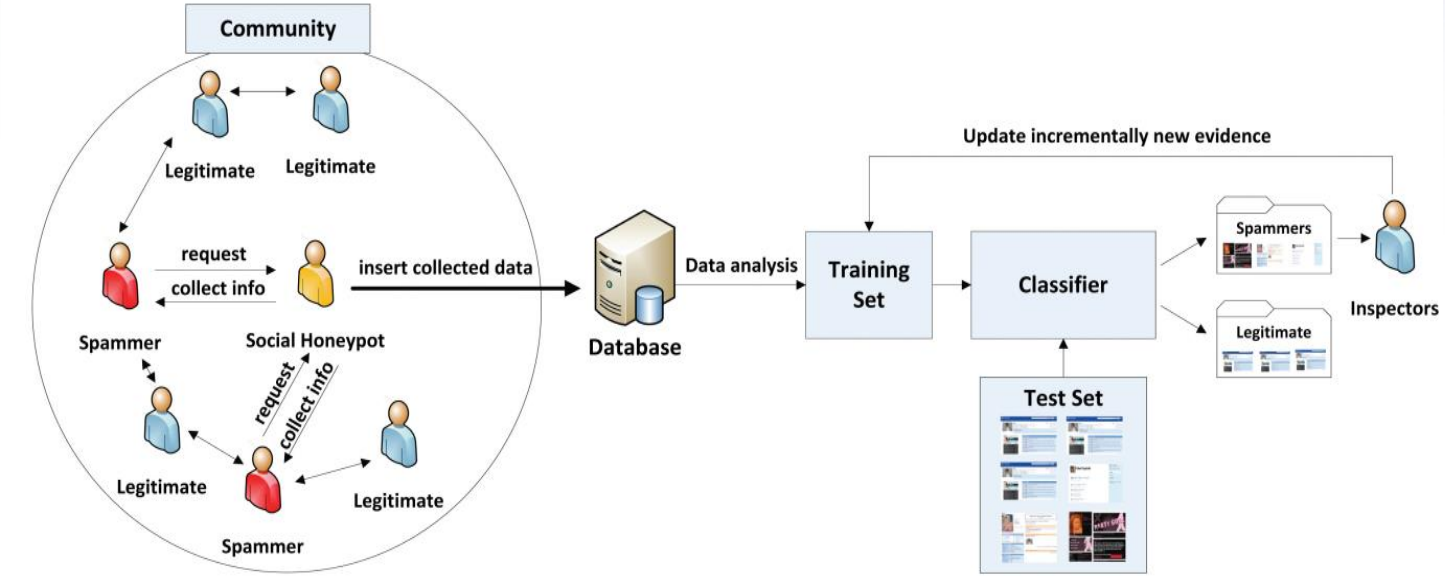
Mitigating Social Spam

- Two defense mechanisms:
  - **Spam content and profile detection**: works at individual level, individual profile content verification etc.
  - **Spam campaign detection**: collective content.

Spam Content and Profile Detection

- Social honeypots are used to capture the behaviour and features of spam profiles.
- The captured features are then used to train a classifier to trap similar spam profiles.
- Social honeypots sometimes mistake legitimate messages.

Social Honeypot



Spammers Types

- From the analysis of the spam profiles' data, the researchers could divide the spammers into four main category:
  - **Displayer** : Spam contents are put on his/her profile, and a friend intentionally visit the profile to check it, and so, it is considered the least effective.
  - **Bragger** : Spammer posts the malicious contents on his own feed, and so, his friends can view it on their feeds.
  - **Poster** : Spammers who send a direct message to each victim, usually by posting on their walls, or sharing the malicious contents in a group or so.
  - **Whisperer** : Those who send private messages to their victims requesting that they, personally and in specific, check a URL to download some files.

Spammers Type

- These 4 categories of spammers can be further divided according to their behavior :
  - **Greedy bot** : When every message contains a malicious content. They are easy to find and flag as spammers.
  - **Stealthy bot** : Usually sends legitimate messages, and include a spam every now and then.

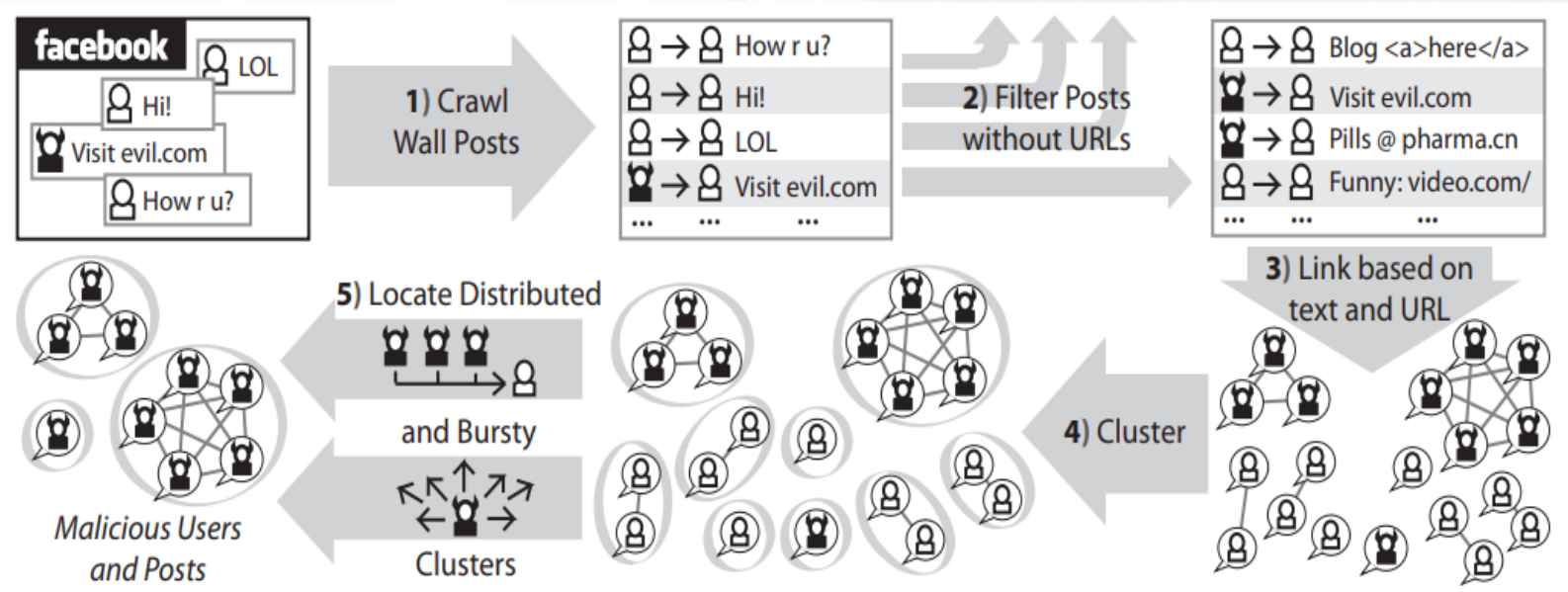
Useful Measures

- **FF ratio** : Compares the number of friend requests that a user sends to the number of friends he has. This is based on the idea that this unknown spammers will have lots of the requests rejected because they do not actually know him.
- **URL ratio** : The high likelihood of a spammer to send out lots of malicious URLs.
- **Message Similarity** : The similarity of messages sent by the spammer to different friends.
- **Number of messages sent** : Based on the observation that spammers tend to send more number of messages compared to legitimate users.

Spam Campaign Detection

- Detecting spam URLs using ML techniques, specifically in twitter tweets.
- Representing post in the form of graphs where each node represents posts.
- Nodes will be connected by an edge if they have same destination URL.
- Spams from same campaign will form subgraphs or clusters.

Spam Campaign Detection



Thanks……