Adaptive Defense Strategies in Social Network

Eshaa Sood, Ashvit Shetty| Network Science (CPSC 8480) | 12-08-2018

Abstract:

YouTube is a video sharing site consisting of various interactive users. Individuals at professional level generate a lot of revenue on the basis of the subscriber counts, number of views and the number of likes. These individuals thus can try and hamper the network in order to make it work in their favor. Also, YouTube being a large social network and users interacting between each other generates tones of personal data which is prone to various malicious attacks. To protect such system from being tampered or attacked our major objective is to develop an analysis model which provides various adaptive defense strategies that can help in protecting the confidentiality of users and avoid the misuse of sensitive data.

Data Set:

YouTube Data set.

This is the data set which is generated on Dec 2008 from YouTube. (http://www.youtube.com/). YouTube is a video sharing site where various interactions occur between users. The data set has 30,522 user profiles which is extracted. For each user, his/her contacts, subscriptions and favorite videos are generated.

Number of Nodes: 15088

Number of Edges: 5574249

Based on the generated information, we construct 5 different interactions between the 15, 088 users. Specifically, they are:

1. The contact network between the 15, 088 users.

2. The number of shared friends between two users in the 848, 003 (excluding the 15,088) contacts.

3. The number of shared subscriptions between two users.

4. The number of shared subscribers between two users.

5. The number of shared favorite videos.

File Description:

There are 6 files included:

1. **nodes.csv:** It's the file of all the users. This file works as a dictionary of all the users in this data set. It's useful for fast reference. It contains all the node ids used in the dataset.

2. **[1-5] edges.csv:** They are the csv format of interactions. Each csv file represents one type of interaction. It is composed of three columns, with the first two representing the user ids, and the last representing the intensity of interaction.

Here is an example:

1,58,3

The interaction intensity between users 1 and 58 is 3. The network is symmetric, so we only show the interaction once. That is, 58, 1,3 will not show up if 1,58,3 is already there.

Method:

Model Description: Our given data set consists of 5 different sets of edges where in the nodes are the users and the edges are the interaction between the various users. The implemented model is an attack defense mechanism on the five networks. On each network we have considered a different attack defense combination and evaluated our defense strategies based on two measures that is Average Clustering and Number of Maximal Cliques.

Average Clustering: The neighborhood of a node, u, is the set of nodes that are connected to u. If every node in the neighborhood of u is connected to every other node in the neighborhood of u, then the neighborhood of u is complete and will have a clustering coefficient of 1. If no nodes in the neighborhood of u are connected, then the clustering coefficient will be 0 [1].

Cliques: A clique, C, in an undirected graph G = (V, E) is a subset of the vertices, $C \subseteq V$, such that every two distinct vertices are adjacent. This is equivalent to the condition that the induced subgraph of G induced by C is a complete graph. In some cases, the term clique may also refer to the subgraph directly [2].

Maximal Cliques: A maximal clique is a clique that cannot be extended by including one more adjacent vertex, that is, a clique which does not exist exclusively within the vertex set of a larger clique. Some authors define cliques in a way that requires them to be maximal and use other terminology for complete subgraphs that are not maximal [2].

Attack and Defense Strategies:

1. **Degree/ Degree Centrality:**
   Degree is a simple centrality measure that counts how many neighbors a node has. If the network is directed, we have two versions of the measure: in-degree is the number of in-coming links, or the number of predecessor nodes; out-degree is the number of out-going links, or the number of successor nodes. A node is important if it has many neighbors, or, in the directed case, if there are many other nodes that link to it, or if it links to many other nodes [3].
2. **Closeness Centrality:**
   In a connected graph, closeness centrality (or closeness) of a node is a measure of centrality in a network, calculated as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Thus, the more central a node is, the closer it is to all other nodes [4].

Time Complexity:

Algorithm	Time Complexity
Degree / Degree Centrality	O(V)
Closeness Centrality	O(V^3)
Average Clustering Coefficient	O(V^3)
No. of Maximal Cliques	O($3^{N/3}$)

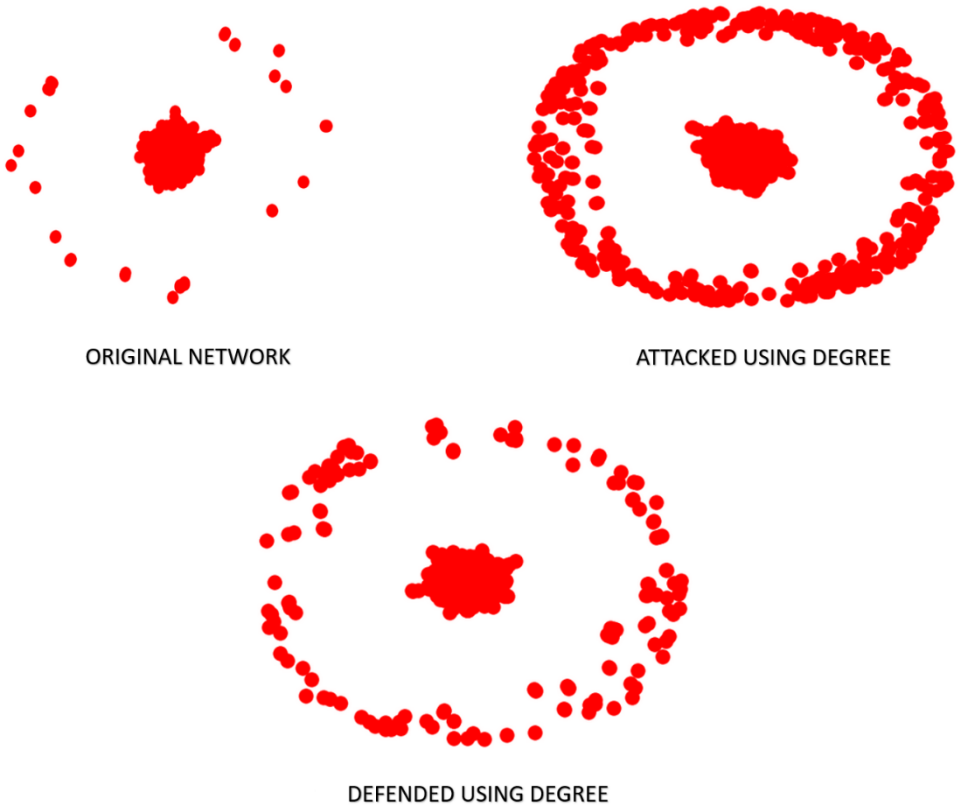
Model Implemented:

- We have designed our own attack defense model based on degree, degree centrality and closeness centrality.
- We have measured the connectivity amongst the network using average clustering and number of maximal cliques.

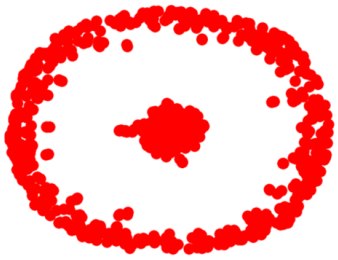
Results:

- ## Network Analysis:

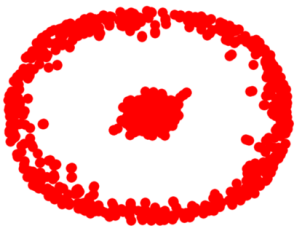
NETWORK 1:



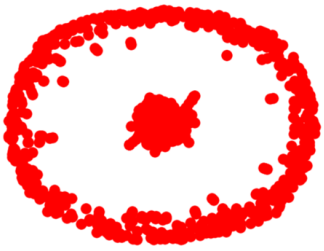
NETWORK 2:



ORIGINAL NETWORK

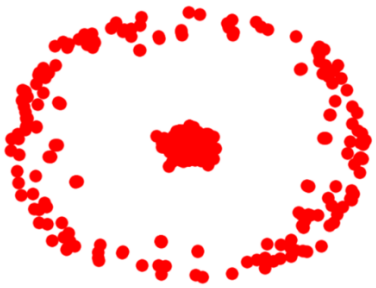


ATTACKED USING DEGREE CENTRALITY

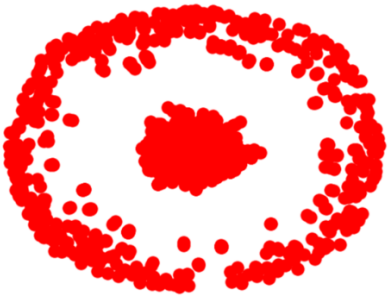


DEFENDED USING DEGREE

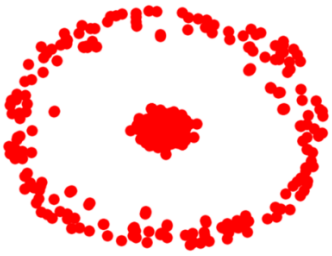
NETWORK 3:



ORIGINAL NETWORK

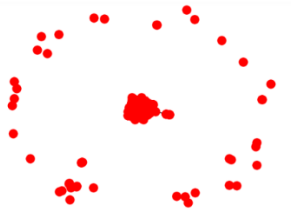


ATTACKED USING DEGREE CENTRALITY



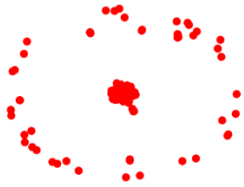
DEFENDED USING DEGREE

NETWORK 4:



ORIGINAL NETWORK

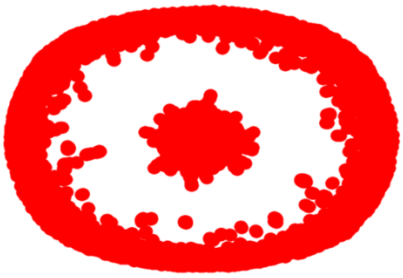
ATTACKED USING DEGREE



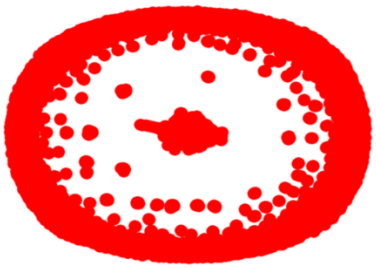
DEFENDED USING DEGREE

NETWORK 5:

ORIGINAL NETWORK



ATTACKED USING DEGREE CENTRALITY



DEFENDED USING CLOSENESS CENTRALITY

- As shown above we have applied different attack defense strategies on each of the five networks and evaluated the efficiency of the defense strategies.
- We can infer from the network plot that the network spreads out after an attack and effective defense strategy would be one that minimizes the spread.
- The defense strategies are **adaptive**, the first four networks have proved to be successful strategies however, the fifth network which was defended by using closeness centrality (protecting neighbors of the important nodes) turned out to be unsuccessful.
- Thus, there is no fixed strategy that would work on all networks. One would have to make a decision based on analyzing past results and trying out various strategies iteratively and come to a conclusion based on drawing comparisons.

- ## Approaches:
  - **Approach 1 (Success):** NETWORK 1

	Original	Attack	Defense
Average Clustering Value	0.1367	0.1237	0.1345
No. of Maximal Clique	69650	49945	60973

Approach 1 was attacked using the degree of value greater than 100 and defended using degree. As the average clustering value and the number of maximal cliques above suggests the defense strategy successfully defended the network.

  - **Approach 2 (Failure):** NETWORK 5

	Original	Attack	Defense
Average Clustering Value	0.1453	0.1297	0.0346
No. of Maximal Clique	9165	7610	2169

Approach 2 was attacked using the degree centrality of value greater than 0.007 and defended using closeness centrality. As the average clustering value and the number of maximal cliques above suggests the defense strategy failed to defend the network.

Similarly, below we have provided with the results from the other networks.

- NETWORK 2

	Original	Attack	Defense
Average Clustering Value	0.3132	0.2941	0.3002
No. of Maximal Clique	63224	35721	36654

- NETWORK 3

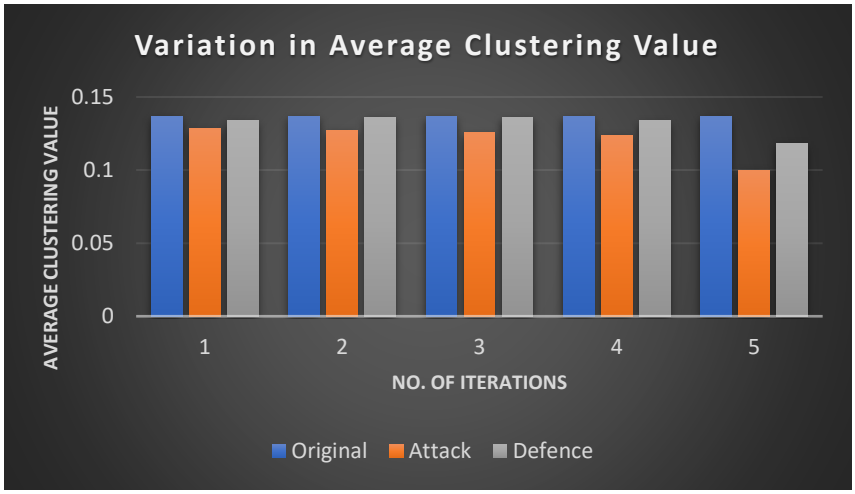
	Original	Attack	Defense
Average Clustering Value	0.6614	0.6187	0.6578
No. of Maximal Clique	27440	2133	10402

- NETWORK 4

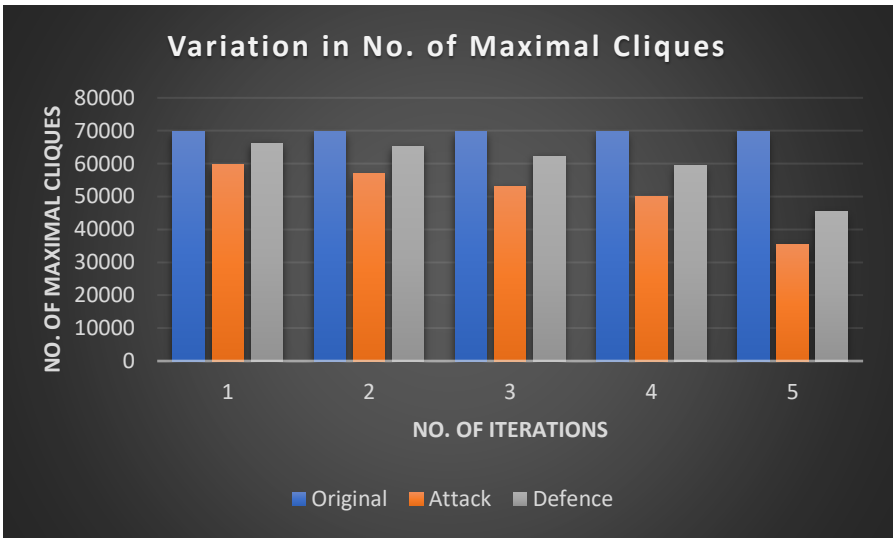
	Original	Attack	Defense
Average Clustering Value	0.6792	0.6779	0.6780
No. of Maximal Clique	130539	127232	127491

Based on the results we can infer that for all the various network plots our analysis model has provided a successful defense.

- **Iterative Evaluation (Network 1):**
  - **Variation in Average Clustering Value.**



  - **Variation in No. of Maximal Cliques**



- Based on our analysis model we have considered two measurement factors that is **Average Clustering value and No. of maximal cliques** to measure the level of connectivity between the users.
- We have evaluated our network by finding the most important node in the network and then iteratively performing the attack and defense strategies.

- For our evaluation we have iteratively increased the intensity of the attack with each iteration.
- As per the results provided above we can see that for every iteration in terms of the number of maximal cliques, the defense mechanism has shown significant results to defend the network.

Conclusion:

Attackers most likely attack nodes with high centrality and thus it is necessary to protect them in order to keep the network connected and the sensitive data intact. We have implemented various defense strategies for the attacks on the YouTube network and evaluated it on two parameters.

References:

[1] Gephi. (n.d.). Gephi/gephi. Retrieved from https://github.com/gephi/gephi/wiki/Average-Clustering-Coefficient

[2] Clique (graph theory). (2018, November 02). Retrieved from https://en.wikipedia.org/wiki/Clique _(graph_theory)

[3] Betweenness Centrality, www.sci.unich.it/~francesc/teaching/network/degree.html.

[4] “Closeness Centrality.” Wikipedia, Wikimedia Foundation, 4 Oct. 2018, en.wikipedia.org/wiki/Closeness_centrality.