TSE Team: int_elligence

Sai Srujan Chinta (sc4401), Ishan Khandelwal (ik2442), Dhruv Chamania (dc3383)

Note: As discussed with Prof Gail in class, we will not be able to present our final demo on May 2nd because Ishan Khandelwal (ik2442) has an exam scheduled on that day (in COMS 4701, AI) and Sai Srujan Chinta (sc4401) has proctoring duties on that day (same course, COMS 4701).

Our project deals with the anonymisation of social networks. The architecture of the model that we have implemented thus far is as follows:

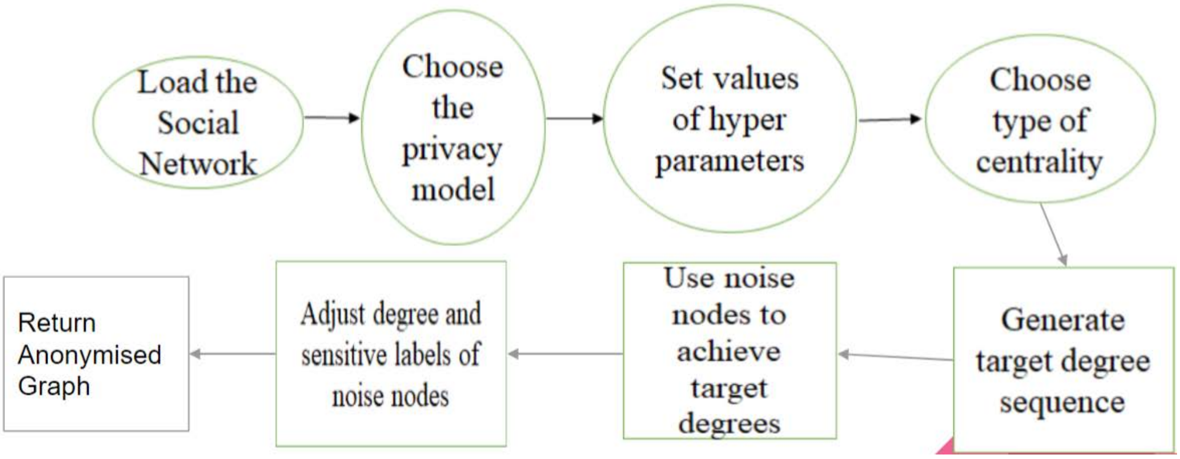
Load the Social Network
Choose the privacy model
Set values of hyper parameters
Choose type of centrality
Generate target degree sequence
Use noise nodes to achieve target degrees
Adjust degree and sensitive labels of noise nodes
Return Anonymised Graph

Figure 1: Architecture of Noise Node Addition Technique

The aim of our project is two fold: Compare and contrast the ease of anonymisation among three possible techniques (edge editing, clustering, noise node addition) of social networks with respect of various parameters, extrapolating the existing literature to account for multiple sensitive attributes. The model architecture above represents our approach to noise node addition technique. The model will remain the same for the other techniques till the “Generate target degree sequence” step. Work done so far: We have implemented the noise node addition technique completely. This entailed implementing two privacy measures (k-degree anonymity and l-diversity), implementing several centrality measures (degree centrality, closeness centrality, eigenvector centrality etc.), generating the target degree sequence (using the K-L based algorithm), using the algorithm suggested in [1] to introduce noise nodes to achieve the target degrees and finally returning the anonymised graph. We have also identified and implemented two structural properties which will be used as parameters to judge the degree of structural change in graphs after anonymising them: Average Path Length and Social Importance. Average Path Length is defined as the sum of the shortest distanes between all pairs of nodes divided by the number of combinations of two nodes. The social importance is characterised by the centrality values.

So far, we have implemented our algorithm on one dataset which comprises data regarding the co-authorship among researchers in the field of networks. The results obtained are as follows:

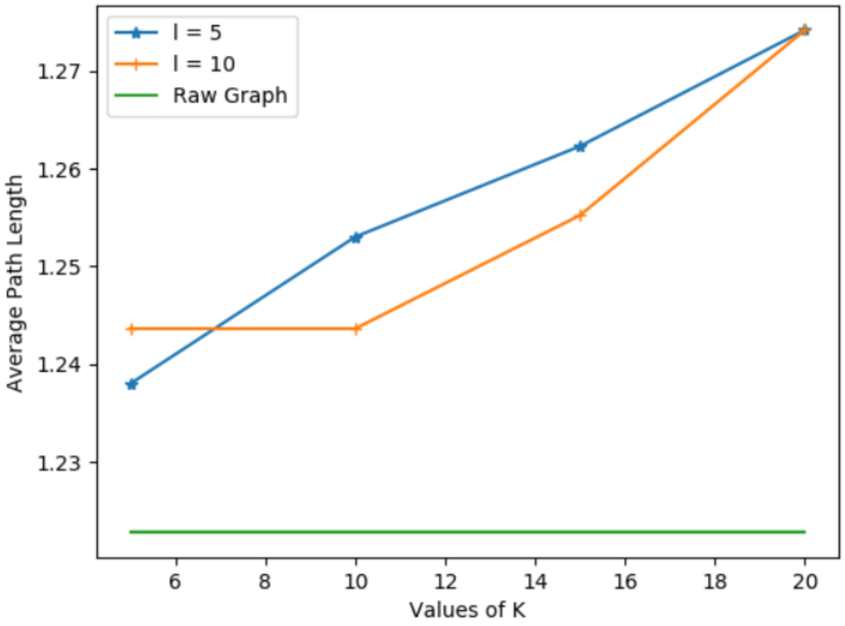

Figure 2: APL of social network with varying k and l

In the above graph, we track the average path length as a function of the hyperparameters k and l. Since the average path length of the original graph is independent of k and l, the green line remains constant.

Where do we go from here?

We will extend our model to include several sensitive attributes instead of just one sensitive attribute. Moreover, after generating the target degree sequence, as mentioned before, there are several ways to achieve the target degree sequence. We will implement other ways and test the ease of developing these methods.

What is new in our project?

To the best of our knowledge, there is no open source code which deals with anonyimising social networks. Moreover, there is very literature regarding extending privacy measures to several sensitive attributes. Best case scenario, we will contribute to the NetworkX library in Python.

Who will use our project?

Any organisation which aims to release graphical data which includes sensitive information will benefit by using our anonymisation model.

What kind of data are we using?

We plan to use the data primarily from the following sources:

1) http://www-personal.umich.edu/~mejn/netdata/ (This data was compiled by Prof Mark Newman)
2) http://snap.stanford.edu/data/#amazon (This data was compiled by Stanford Labs)
3) https://github.com/gephi/gephi/wiki/Datasets

Research Questions that we plan to answer:

1) How do the structural properties of social networks change as a consequence of being subjected to anonymisation techniques?
2) From a software development perspective (analysing time and space complexity), how feasible is it to implement some of the anonymisation techniques?
3) How can we extend the definiton of the privacy measures to account for multiple sensitive attributes without compromising the structural integrity of the original social network?

We will share the code and documentation via a private Github repository. We will give the instructors access to our repository in the near future.

We are splitting the work equally among the team members. However, more specifically, Sai Srujan and Ishan will primarly focus on developing the new algorithms and main responsibilty for Dhruv will be to focus on the testing aspect.