Project Proposal

Research Question

Reducing/optimizing classification error rate for classification problems in sparsely distributed graphs using graph based semi supervised learning. Due to the presence of limited labelled data in the real world, semi-supervised learning approaches are an efficient way to compound unsupervised learning methods with available data and use them to the best potential. The models we plan on optimizing are the following

Method

We plan to carry out our optimisation by considering the following models

- Label Propagation
- Shoestring
- Deep Walk
- Node2Vec

and tweaking their parameters and making graph structure assumptions. Then applying those tweaked models to several datasets, we will compare classification accuracy of the naive method against our own tweaked method.

Why this problem?

This is a worthwhile investigation due to the difficulty in obtaining labeled data sets in the real world, and the expense in manually labeling data. It is often much more convenient to obtain partially labeled data, and there are feasible models that can apply to these datasets and perform nearly as well as supervised learning.

Datasets we plan to use

5-7 datasets form **Stanford Large Network Dataset Collection**

We also plan to use datasets in the realm of citation networks, social networks and judicial rulings as a few real world cases of how semi-supervised learning methods can be made more efficient and solve actual issues in the process.

Members - Ridhit Bhura (rb749), Ishaan Chansarkar (ic254), Kenan Clarke (kc676)