Data mining course proposal

Lucas….
Maryam Kheirkhahzadeh markhe@kth.se
Pablo Laso plaso@kth.se

1. Introduction

In network theory, link prediction is the problem of predicting the existence of a link between two entities in a network. In this project, we aim to compare the likeness of ML algorithms to that of Edge prediction models, on online social networks graphs and five collaboration networks.

2. Research path (which algorithms and model we want to use):

We would like to find the similarity between some machine learning models and some different edge prediction approaches, which we have introduced during the course and from reading related papers. We call them class1 and class2 approaches. The machine learning models we have chosen for class1, and different edge prediction approaches for class2, are as follows:

- **Class1:Machine-learning models**: Logistic regression, Decision tree, Support vector machines, Lasso regression, Random forest.

- **Class2:Edge prediction policies:** Triadic closure, Jaccard Coefficient, Resource Allocation Index, Adamic Adar Index, Preferential Attachment. We should use the Networkx library to implement these methods, except the first one. There is no built-in function for the Triadic closure.

1) **Triadic closure**: If two vertices are connected to the same third vertices, the tendency for them to share a connection is Triadic Closure. Example: if there is edge (1,2) and (2,3) then, the Triadic closure predicts edge (1,3).
2) **Jaccard coefficient**:It is calculated by number of common neighbors normalized by total number of neighbors. We will use the pre-defined function jaccard_coefficient. It returns a list of 3 tuples (u, v, p), where u, v is the new edge which will be added next with a probability measure of p (p is the Jaccard Coefficient of nodes u and v).
3) **Resource Allocation Index**:It is defined as a fraction of a resource that a node can send to another through their common neighbors. We will use the built-in function resource_allocation_index which offers a list of 3 tuples

(u, v, p), where u, v is the new edge and p is the resource allocation index of the new edge u, v.

4) **Adamic Adar:**The built-in function adamic_adar_index returns a list of 3 tuples (u, v, p) where u, v is the new edge and p is the adamic adar index of the new edge u, v.
5) **Preferential Attachment:**Preferential attachment means that the more connected a node is, the more likely it is to receive new links. The built-in function in the networkx package is preferential_attachment.

3. **Measures to compare two classes:(How will you evaluate your method? How will you test it? How will you measure success?)**

   We want to compare two classes with AUC, accuracy, Brier score, precision and recall.

4. **Data**

   We want to test and implement our idea on three directed social network graphs and 5 collaboration networks. We can then compare these two types of networks.

   Three social network graphs:

- The *Epinions* dataset. Node i trust on node j. This is the link to the dataset:https://snap.stanford.edu/data/soc-Epinions1.html
- Wikipedia vote network: directed graph. Node i voted for node j. Thi is the link to the dataset:https://snap.stanford.edu/data/wiki-Vote.html
- Directed LiveJournal friendship social network, This is the link to the dataset:https://snap.stanford.edu/data/soc-LiveJournal1.html

Five collaboration Networks are:

- Collaboration network of Arxiv Astro Physics. The dataset link is: https://snap.stanford.edu/data/ca-AstroPh.html
- Collaboration network of Arxiv Condensed Matter. The dataset link is: https://snap.stanford.edu/data/ca-CondMat.html
- Collaboration network of Arxiv General Relativity. The dataset link is: https://snap.stanford.edu/data/ca-GrQc.html
- Collaboration network of Arxiv High Energy Physics: https://snap.stanford.edu/data/ca-HepPh.html
- Collaboration network of Arxiv High Energy Physics Theory: https://snap.stanford.edu/data/ca-HepTh.html

5. The problem we want to solve

We try to compare some machine learning algorithms (class1) with some edge prediction approaches (class2) on two kinds of networks. We have decided to implement and test our idea on three online social networks graphs and five collaboration networks and try to find out if there is any similarity between these two classes for link prediction or not. Likewise, we can see that many papers use these methods to predict links in networks, so understanding the similarity or differences between measures could help researchers to select better approaches in the link prediction problems. Finally, we will report the result based on different evaluation metrics (AUC, Accuracy, etc).

6. What do we expect to submit/accomplish by the end of the quarter

We expect to obtain reliable results on link prediction by using both ML models and Edge Detection algorithms. Furthermore, we expect to submit a grounded discussion on the comparison between ML model and Edge Detection algorithms when aiming at predicting links in networks.