# EECS4414/5414 Winter 2024

# Information Networks

# Assignment 2 (10%): Link Analysis & Prediction; Community Detection

**Posted**: Fri, May 17, 2024; **Due**: 11:59 pm on Fri, Mar 31, 2024

## Objective

In this assignment, you will be writing programs for generating temporal snapshots of a network and performing basic link analysis of the network. Then, you will work on methods for link prediction in these networks. Furthermore, you will experiment with community detection methods. You will also prepare a technical document where you briefly report about the data, methods, measurements, experiments, and results of the assignment.

Important Notes:

- If you are working in a pair, **only one of you should submit**. The first page of your report should clearly include information of all team members (first name, last name, login, student#, email).
- You can write code of your own or use available graph libraries (NetworkX, SNAP, JUNG, etc.).
- Your report should be **no more than 5 pages**. To get full marks, your code must be well-documented, and the report should be well organized, using proper technical language and suitable style (sections, figures, tables).

## Formatting and Style

All reports should be formatted according to the ACM SIG conference proceedings template in LaTex and prepared using Overleaf, a free collaborative authoring tool. The template can be accessed here: https://www.overleaf.com/latex/templates/association-for-computing-machinery-acm-sig-proceedingstemplate/bmvfhcdnxfty

## Electronic Submission Instructions

You should submit your work electronically using the `submit` command in PRISM lab computers. For this assignment, you will submit two files:

- your *code* (**a2-code.zip**), and
- your *report* (**a2-report.pdf**)

When you have completed the assignment, move these two files in a directory (e.g., assignment1/), and use the following command within that directory to electronically submit your files:

```
% submit 4414 a2 a2-code.zip a2-report.pdf
```

You may submit your files as many times as you wish prior to the submission deadline. Make sure you name your files exactly as stated (including lower/upper case letters). You may check the status of your submission using the command:

```
% submit -l 4414 a2
```

- Make sure you have submitted the correct version; new or missing files will not be accepted after the due date.

## A. Temporal Graphs (20%)

Consider the dataset "**dblp_coauthorship.json.gz**" available at the following website:
http://projects.csail.mit.edu/dnd/DBLP/
This dataset represents a DBLP co-authorship graph consisting of (author1, author2, year) triples. Based on this dataset, create the following graphs:

(a) dblp2005: An undirected unweighted graph that represents the DBLP co-authorships of the year 2005. Each node represents an author, and each edge represents that two authors co-authored at least one paper in 2005.
(b) dblp2006: An undirected unweighted graph that represents the DBLP co-authorships of the year 2006. Each node represents an author, and each edge represents that two authors co-authored at least one paper in 2006.
(c) dblp2005w: A weighted version of the dblp2005 graph, where the weight represents the number of papers that two authors co-authored in 2005.

For all graphs above, you need to obtain and work only on the giant connected component. Report the number of nodes and edges of the graphs (#nodes, #edges).

## B. Node and Edge Importance in Graphs (30%)

For each of the graphs above (dblp2005, dblp2006, dblp2005w) report:

i. the names and scores of the 50 most important authors based on PageRank scores
ii. the pairs of author names of the 20 most important edges based on edge betweenness scores

Briefly comment on the results and how they compare with each other.
Use small text font to fit the results in the report.

## C. Link Prediction in Graphs (40%)

Given the co-authorship network of 2005, we want to predict the new co-authorships that will occur in 2006 (i.e., the new edges). Your prediction will be based on a number of prediction methods proposed in the literature and your task is to evaluate their performance. The steps to follow are described below:

i. based on dblp2005 create a graph dblp2005-core that includes nodes with degree d>=3.
ii. based on dblp2006 create a graph dblp2006-core that includes nodes with degree d>=3.
iii. compute the list of friends-of-friends FoF in dblp2005-core; this is the list of pairs of nodes that are exactly two-hops away in the network. FoF is the list of candidate edges to consider for the prediction problem.
iv. compute the set of edges T that do not exist in dblp2005-core but exist in dblp2006-core. This is the set of target edges that we would ideally be able to predict.
v. compute the set of predicted edges P according to the following link prediction methods:
   a. RD: random predictor
   b. CN: common neighbors
   c. JC: jaccard coefficient
   d. PA: preferential attachment
   e. AA: adamic/adar
vi. compute the precision at k evaluation metric for values of k = {10, 20, 50, 100, |T|}, denoted as P@10, P@20, P@50, P@100, P@T for each method.

Briefly comment on the results and how they compare with each other.

## D. Community Detection in Graphs (10%)

Pick the co-authorship network of 2005 and apply the Girvan–Newman community detection method to derive communities. The Girvan-Newman method iteratively removes the edge with the highest edge betweenness to derive communities (hierarchical method that creates a dendrogram). Stop the iterative process once the current number of communities is greater than `k=10`. Report the size (number of nodes) of the 10 (or more) communities currently found in a descending order (first the size of the largest community and last the size of the smallest community).

Briefly comment on the results.