CS 6402 – Advanced Topics in Data Mining

Course Project

This assignment will be worth **30%** of your course grade.

This assignment is due by **11:59 p.m. on Friday, May 1st**

Basic Instructions

Working together with **at most 1 other** student enrolled in this course, you are to complete some original project that is focused on **graph data mining**. This is expected to be some combination of design, implementation, experimentation, analysis, and/or written reporting components.

Projects fall into **two categories** depending on what the primary "deliverable" is: **(1) software** and **(2) research papers**. A **software project** is expected to involve extensive programming; however, professional-style written documentation also will be required (i.e., a short write-up of technical documentation, description of datasets tested, experimental/benchmark results, etc.). A **research paper** could include some programming (e.g., as proof-of-concept for some short algorithm(s), implementations of existing algorithms for testing, etc.); however, the primary deliverable is expected to be an 8-10 page formal, "research-style" paper (including sections for an introduction, background/related work, methods, conclusions, etc.)[1].

Submission Requirements

Only **one** person on a team **should submit** the project for grading. A project can be submitted any time before the due date, and can be submitted multiple times; only your last submission will be graded. Unless otherwise noted below, any files you need to submit for your project should be submitted electronically via **Canvas**; a **Project** submission site for doing so will be set up under ***Assignments***.

Contributions of Each Team Member

Regardless of what kind of project you do, you must write one paragraph that summarizes **your** primary contributions to the project (e.g., I performed all the research for (and wrote) the Background and Related Work section of the paper, I did most of the algorithm development, etc.). Be specific and honest! This will, in part, determine your grade on this assignment.

You **ALSO** must write one paragraph that summarizes the primary contributions of **the other person** who worked on the project (e.g., John designed and ran the experiments, Jane analyzed the experimental data and actually wrote the majority of the paper, etc.). Again, be specific and honest!

Create a **single *pdf* file** containing these two paragraphs, and name the file *eval_* plus your last name followed by your first initial (e.g., John Smith would name his file *eval_smithj.pdf*). **EMAIL** this *pdf* file to Dr. Leopold (leopoldj@mst.edu) by 5 p.m. on the project due date; do **NOT** submit this file via Canvas! **Each person must submit this; otherwise, s/he will not receive a**

[1] Papers must be in IEEE standard double column format:
https://www.ieee.org/conferences_events/conferences/publishing/templates.html

grade for his/her project! Of course, if you do the project by yourself, you are exempt from this requirement.

Software

If you do a software project, then you must make the documentation, source code, and test files available by providing Dr. Leopold with a zip file by the project due date. **Note**: If Dr. Leopold **can't compile and execute** your software, you will receive a grade of **0** for this project; you will **NOT** be given an opportunity to resubmit if she informs you that she can't compile or execute your program! For that reason, in addition to providing the software deliverables, you also may want to schedule a **demonstration** of your software **at least one week** before the project due date.

Papers

If you do a research paper, you should create a **single *pdf* file** for your paper and name it using the combination of the last names of the people who worked on the project (e.g., if John Smith and Jane Doe worked together, their submission should be named *SmithDoe.pdf* or *DoeSmith.pdf*); if you write the paper by yourself, just name the file using your last name and first initial. **Note**: Papers that are not properly formatted will **not** be graded; you will **NOT** be given an opportunity (after submission) to go back and reformat your paper if Dr. Leopold informs you that your paper is in the wrong format!

Extra Credit!

You can earn **5% extra credit** by giving a **20-minute presentation** about your project at the end of the semester. If you are interested in doing this, you will need **to notify Dr. Leopold by April 24, 2020** so that your presentation can be worked into the lecture schedule.

Some Project Ideas

Software Projects

- Take an existing FSM algorithm that works for a transaction graph setting, and implement a version that works for a single graph; benchmark it on a variety of real and/or synthetic graphs
- Modify an existing GDM algorithm to work via Hadoop or Spark; benchmark it on a variety of real and/or synthetic graphs (including fairly big graphs); if possible, benchmark your implementation against a non-distributed implementation of the algorithm
- Do the same as above, but using parallelization
- Significantly modify (or write a completely new!) GDM algorithm and implement it; benchmark it against other similar algorithms or against a "base" algorithm using a variety of real and/or synthetic graphs to show that the modification is an improvement in terms of computation time and/or memory usage

Research Papers

- Something associated with your M.S. or Ph.D. research that involves graph data mining
- A novel application of existing GDM method(s) to a dataset; note that this will require applying existing software for the GDM method(s)
- A survey on a topic in graph data mining that we're not going to cover this semester; must read 6-8 papers on the topic (not including papers we read in class); focus should be on **algorithms/methods**, **not applications**!

Warning: Before you begin working on something, do a search for it on the internet to see whether it has already been done. You will NOT get credit for something that is not novel!